Journal of Clinical Epidemiology 105 (2019) 101−111

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Two valid and reliable work role functioning questionnaire short versions were developed: WRFQ 5 and WRFQ 10

Femke Abma[a,*], Jakob Bue Bjorner[b,c,d], Benjamin C. Amick III[e,f], Ute Bültmann[a]

[a]*University of Groningen, University Medical Center Groningen, Department of Health Sciences, Community and Occupational Medicine, Groningen, the Netherlands*

[b]*National Research Centre for the Working Environment, Copenhagen, Denmark*

[c]*Optum Patients Insights, Johnston, RI, USA*

[d]*Department of Public Health, University of Copenhagen, Copenhagen, Denmark*

[e]*Florida International University, Robert Stempel College of Public Health & Social Work, Miami, FL, USA*

[f]*Institute for Work & Health, Toronto, Ontario, Canada*

Accepted 18 September 2018; Published online 22 September 2018

Abstract

Objective: The study aims to develop and validate short versions of the work role functioning questionnaire v2.0 (WRFQ) that retain the measurement properties of the full-length 27-item questionnaire.

Study design and setting: Six cross-sectional Dutch samples (N = 2,433) were used, containing data on gender, self-rated health, job type, and WRFQ scores. Indicators from classical test theory and item response theory methods were used along with evaluation of translatability and conceptual considerations to identify short version candidate items. To ensure content validity, the item selection was made within the five-factor structure established for the WRFQ—leading to a 5-item and a 10-item short version. Bland-Altman analyses of agreement and interclass correlations with the full WRFQ were used to establish the best scoring procedure. Discriminant validity was evaluated for the short versions and compared with the full-length 27-item version.

Results: Both short versions showed acceptable agreement with the full-length 27-item version using simple scoring procedures. Both also showed comparable or stronger validity than the full WRFQ in known group comparisons.

Conclusion: Both short versions can be used to measure work role functioning in working samples with mixed clinical conditions and job types.

Keywords: Employee health; Work performance; Validation; Measurement; IRT; Lost productivity

1. Introduction

Over the past 25 years, a number of self-report questionnaires have emerged to assess lost productivity at work [1−6]. These questionnaires include the Stanford presenteeism scale [7], the work limitations questionnaire, [8] and the health and productivity questionnaire [9] to mention a few. For some, in addition to a full-length version, one or more short versions exist. All questionnaires aim to measure the intersection of a person's health and work performance. These questionnaires have been frequently used in clinical trials to assess work-related outcomes of medical interventions [10]. The information can be used in provider-patient interactions. More recently, these questionnaires have also been used to measure work performance after return to work, e.g., after sickness absence because of common mental disorders or cancer [11,12]. The majority of the questionnaires were developed in the 20th century, based on 20th century models of work. Yet today in a new world of work with changing workplaces, work practices, and technologies, new challenges arise [13,14].

To address the changing nature of work, version 2.0 of the work role functioning questionnaire (WRFQ) was developed [15]. The WRFQ v2.0 builds on an earlier questionnaire but adds a new dimension, flexibility demands, reflecting the 21st century workplace. To the best of our knowledge, no existing health-related work questionnaire includes these new flexibility demands. The WRFQ v2.0 measures the perceived difficulties in meeting work

Conflict of interest: None.

* Corresponding author. University of Groningen, University Medical Center Groningen, Department of Health Sciences, Community and Occupational Medicine, Groningen, the Netherlands. Tel.: +3150 3615081; fax: +3150 3636251.

E-mail address: f.i.abma@umcg.nl (F. Abma).

https://doi.org/10.1016/j.jclinepi.2018.09.005

What is new?

Key findings

- Both short versions reflect the five-factor structure of the full-length 27-item version, have acceptable agreement with the full-length 27-item version, and showed acceptable measurement properties.

What this adds to what was known?

- This article presents the development of two short versions of the WRFQ v2.0, a 5- and a 10-item version.

What is the implication and what should change now?

- Both versions can be used to measure work role functioning in working samples with mixed clinical conditions and job types.
- The choice between the 5-, 10-, and 27-item versions depends on the intended use of the instrument and is a compromise between length and measurement properties.

demands among employees given their physical health or emotional problems [16−19]. The WRFQ v2.0 consists of 27 items, summarized in four subscales (work scheduling and output demands, physical demands, mental and social demands, and flexibility demands) or in a total score. Recent confirmatory factor analyses (CFA) of six samples found support for a five-factor structure representing five domains (separating work scheduling from output demands) [20]. For several items, the analyses revealed potential redundancies, suggesting that a shorter version of the questionnaire might be developed.

This study aims to develop and evaluate shorter versions of the WRFQ v2.0 retaining the measurement properties of the full-length questionnaire. Such shorter versions may help clinicians and other practitioners to start a conversation with the patient/employee, followed by the use of the full-length version to explore the perceived difficulties at work in more detail. Shorter versions may also be preferable in research with limited questionnaire space. Shorter versions of the WRFQ v2.0 should reflect the measurement properties of the full-length version, e.g., reliability and validity, and permit calculation of scores comparable to the full-length version. The two aims of this study are 1) to select items to develop two short versions of the WRFQ 2.0; and 2) to validate the short versions. In the validation, we will evaluate the ability of the two short versions to reproduce the total score of the full-length 27-item WRFQ v2.0, compare their measurement properties and ability to discriminate known groups with the full-length version.

2. Methods

2.1. Work role functioning questionnaire v2.0

The WRFQ measures the perceived difficulties in meeting work demands among employees given their physical health or emotional problems [15,16]. The WRFQ was administered using a 4-week recall period and six response options: 0 = difficult all the time (100%), 1 = difficult most of the time, 2 = difficult half of the time (50%), 3 = difficult some of the time, 4 = difficult none of the time (0%). The sixth response option "Does not apply to my job" was included to allow a respondent to validly answer when the work demand was not part of the job. It was coded as missing. The scoring of the subscales and the total score used a simple summative approach, taking the average of the items multiplied by 25 to obtain scores between 0 and 100, with higher scores indicating better work functioning. If more than 20% of the items were missing, the scale score was set to missing. Based on the results of previous CFA [20], some items were flagged as problematic: item 9 (feel a sense of accomplishment) showed local correlation with item 10 (feel you have done what you are capable of doing). Similarly, item 18 (concentrate on your work) showed local correlation with item 19 (work without losing train of thought). Finally, item 15 (use hand-held tools), which is hypothesized to be part of physical demands, showed cross-loadings with the mental and social demands domains.

2.2. Study samples

Table 1 shows the six cross-sectional samples (N = 2,433) used for the investigation. The samples were collected from various populations in the Netherlands between 2010 and 2014 and described in more detail elsewhere [20]:

1. A sample from the general working population (*general working population*) [15], i.e., a heterogeneous sample of workers across job types and health status;
2. A sample of shift workers (*shift worker population*) [21] with a regular shift, shift workers with irregular shifts, on call workers, and workers on day shifts. Regarding health status this is a heterogeneous sample;
3. A sample of employees diagnosed with cancer, which returned to work in the last 3 months for at least 12 hours per week (*cancer diagnosis population*) [22]. The sample is heterogeneous with respect to job type and cancer diagnoses (e.g., breast cancer, gastrointestinal cancer, gynecological cancer, hematological cancer, urogenital cancer);
4. A sample of occupational and insurance physicians (*occupational and insurance physicians population*). This sample was asked to complete the questionnaire when attending a 1-day conference and is rather homogeneous.
5. A sample of university workers (*university worker population*) [23], heterogeneous regarding job type (both academics and supporting staff) and health status;

Table 1. Description of the six samples

N	General workers *N* = 553	Shift workers *N* = 1,055	Cancer patients *N* = 229	Occupational and insurance physicians *N* = 154	University workers *N* = 284	Common mental disorder patients *N* = 158
Age, M(SD)	45.1 (10.6)	44.0 (10.1)	50.8 (7.9)	53.7 (6.2)	45.6 (10.9)	42.3 (9.6)
Gender, N (%)						
Male	338 (70.2)	922 (87.4)	91 (1.3)	93 (60.4)	125 (44.0)	65 (41.1)
Female	165 (29.8)	117 (11.1)	135 (59.0)	52 (33.8)	159 (56.0)	93 (58.9)
Job Type, N (%)						
Manual	156 (28.2)	256 (24.3)	23 (1.3)	0 (0)	[a]	[a]
Nonmanual	257 (46.5)	91 (8.6)	139 (60.7)	145 (100)	110 (61.3)	[a]
Mixed	5 (0.9)	638 (60.5)	64 (27.9)	0 (0)	[a]	[a]
Health status, N (%)						
Excellent-very good-good	491 (88.8)	883 (83.7)	170 (74.2)	134 (87.0)	248 (87.3)	110 (69.6)
Fair/poor	58 (10.5)	95 (9.0)	56 (24.5)	11 (7.1)	35 (12.3)	45 (28.5)
Work role functioning total score, M (SD)	84.2 (15.8)	86.9 (13.7)	77.3 (17.6)	83.0 (12.6)	84.8 (14.4)	[b]
Work scheduling demands	83.0 (21.7)	86.6 (17.6)	77.3 (21.3)	80.9 (22.1)	83.9 (19.8)	65.3 (24.3)
Work output demands	81.0 (20.9)	84.7 (18.0)	74.6 (23.0)	76.6 (16.7)	79.8 (20.5)	64.7 (23.7)
Physical demands	87.1 (19.6)	89.0 (16.9)	83.7 (19.3)	94.0 (13.2)	91.6 (15.6)	90.5 (21.9)
Mental and social demands	85.2 (17.5)	87.5 (15.3)	75.4 (21.2)	85.8 (12.9)	85.0 (15.6)	64.1 (20.5)
Flexibility demands	84.0 (20.7)	87.4 (15.8)	78.4 (20.9)	80.3 (16.1)	85.1 (16.8)	[b]
Chronic conditions						
0	160 (28.9)	[a]	112 (48.9)	[a]	[a]	[a]
1	72 (13.0)	[a]	66 (28.8)	[a]	[a]	[a]
2	23 (4.2)	[a]	32 (14.0)	[a]	[a]	[a]
3 or more	18 (3.3)	[a]	19 (8.3)	[a]	[a]	[a]

Numbers might be lower/not add up to 100% due to missing.
[a] No information available.
[b] Flexibility demands items missing, therefore no comparison score available.

6. A sample of workers who had partially or fully returned to work 3 months after a period of sick leave because of common mental disorders (*common mental disorder population*) [11]. The sample is heterogeneous regarding job type and contains workers with various common mental disorders (e.g., adjustment disorders, anxiety disorders, mild depression).

2.3. Measures

For each sample, the following information was available:

- Gender (male/female)
- Self-rated health (excellent, very good, good/fair, poor) measured with the first question of the SF12 [24].
- Job type (manual/nonmanual/mixed, except in the university worker population, which only distinguishes between university vs. supporting staff)
- WRFQ v2.0 (except the common mental disorder population, which did not contain the flexibility demand items because data were collected before the development of these items)

In addition, in the general working population and common mental disorder population samples, information about the number of chronic conditions was available.

2.4. Statistical analyses

The analyses were conducted in four steps: 1) psychometric analyses at item level, 2) definition of two short versions, 3) developing procedures to map the short versions to the total score of the full-length 27-item WRFQ 2.0, and 4) scale level psychometric analyses.

2.5. Item-level statistical analyses

Descriptive analyses, analyses of differential item functioning (DIF), and analyses based on item response theory (IRT) were used to explore statistical properties of the items. Criteria for problems in the descriptive analyses were items with more than 15% missing and items with more than 50% of responses at best category (ceiling). Items below these cutoffs obtained a score of 0 (good), items exceeding this cutoff were scored with 1 (poor). Tests of DIF were conducted with regard to the population (each data set), gender, age, and self-rated

Table 2. IRT parameter estimates—Graded Response Model

	Discrimination		Difficulty 1		Difficulty 2		Difficulty 3		Difficulty 4		Item fit	
Subscale/Item	Est	*SE*	Est	*SE*	Est	*SE*	Est	*SE*	Est	*SE*	S-G^2	*P*
Work scheduling demands												
WRFQ1	2.81	0.16	−1.73	0.08	−1.38	0.07	−0.19	0.04			61.43	0.0000
WRFQ2	4.29	0.26	−1.70	0.06	−1.32	0.05	−0.44	0.03			29.01	0.0344
WRFQ3	1.84	0.11	−2.10	0.12	−1.45	0.08	−0.22	0.05			58.96	0.0000
WRFQ4	1.78	0.12	−2.12	0.13	−1.59	0.09	−0.47	0.05			40.40	0.0067
Output demands												
WRFQ5	3.09	0.17	−2.89	0.15	−2.00	0.09	−1.37	0.06	−0.32	0.03	79.14	0.0000
WRFQ6	2.92	0.15	−2.75	0.15	−1.86	0.09	−1.16	0.05	0.03	0.03	115.65	0.0000
WRFQ7	3.21	0.16	−2.73	0.13	−1.88	0.08	−1.48	0.07	−0.15	0.03	96.36	0.0000
WRFQ8	3.12	0.17	−2.52	0.12	−1.76	0.08	−1.27	0.06	−0.32	0.03	53.85	0.0124
WRFQ9	1.62	0.10	−3.05	0.20	−1.96	0.11	−1.22	0.08	0.04	0.05	75.04	0.0005
WRFQ10	1.97	0.12	−2.54	0.14	−1.69	0.09	−1.32	0.08	−0.31	0.05	75.85	0.0004
Physical demands												
WRFQ11	2.75	0.20	−2.50	0.17	−1.93	0.10	−1.63	0.08	−0.58	0.05	15.60	0.7410
WRFQ12	3.06	0.20	−2.51	0.16	−1.80	0.10	−1.28	0.07	−0.38	0.04	20.79	0.5335
WRFQ13	4.55	0.30	−2.25	0.10	−1.70	0.07	−1.20	0.05	−0.34	0.03	12.07	0.7961
WRFQ14	5.61	0.38	−2.15	0.09	−1.66	0.07	−1.22	0.04	−0.35	0.03	27.02	0.0577
WRFQ15	2.37	0.21	−3.27	0.32	−2.41	0.15	−1.84	0.11	−1.00	0.06	26.01	0.1298
Mental and social demands												
WRFQ16	3.74	0.20	−2.75	0.15	−2.16	0.09	−1.50	0.06	−0.04	0.03	37.28	0.0411
WRFQ17	3.69	0.21	−2.74	0.14	−2.09	0.11	−1.63	0.07	−0.41	0.03	52.48	0.0007
WRFQ18	6.51	0.43	−2.64	0.09	−1.93	0.07	−1.37	0.04	−0.08	0.02	21.66	0.4194
WRFQ19	4.54	0.26	−2.82	0.14	−2.02	0.08	−1.33	0.05	0.00	0.03	45.50	0.0035
WRFQ20	2.74	0.14	−3.11	0.19	−2.12	0.10	−1.39	0.06	0.03	0.04	39.10	0.0621
WRFQ21	1.85	0.12	−3.72	0.30	−2.77	0.17	−2.07	0.11	−0.77	0.06	88.12	0.0000
WRFQ22	1.74	0.12	−3.62	0.26	−2.74	0.17	−2.19	0.13	−0.80	0.06	62.87	0.0001
Flexibility demands												
WRFQ23	2.99	0.19	−2.69	0.18	−1.99	0.11	−1.45	0.07	−0.26	0.04	35.44	0.0123
WRFQ24	2.55	0.17	−2.87	0.23	−2.19	0.14	−1.61	0.08	−0.35	0.04	34.11	0.0178
WRFQ25	2.27	0.14	−2.91	0.23	−1.98	0.12	−1.24	0.07	−0.01	0.05	37.34	0.0299
WRFQ26	4.02	0.26	−2.70	0.17	−1.78	0.08	−1.16	0.05	−0.08	0.03	10.29	0.8908
WRFQ27	2.97	0.21	−2.62	0.18	−2.08	0.13	−1.65	0.09	−0.74	0.04	42.02	0.0011

For the items concerning work scheduling demands: Difficulty 1, threshold for answering "difficult half of the time" or better; Difficulty 2, threshold for answering "difficult some of the time" or better; Difficulty 3, threshold for answering "difficult none of the time."

For the items concerning output demands, physical demands, mental and social demands, and flexibility demands: Difficulty 1, Threshold for answering "difficult most of the time" or better; Difficulty 2, Threshold for answering "difficult half of the time" or better; Difficulty 3, Threshold for answering "difficult some of the time" or better; Difficulty 4, Threshold for answering "difficult none of the time."

health using a logistic regression approach [25]. Important DIF was identified by statistical significance and a Nagelkerke R^2 difference larger than 2%. All items were ranked based on the R^2 difference score, with lower scores indicating lack of DIF and higher scores indicating more severe DIF.

Preliminary IRT analyses were performed in a data set that combined the two largest samples: General workers and shift workers, because DIF analyses had found very little DIF between these two samples (results provided by first author on request). The IRT analyses were done separately for each subscale to identify the best items for each domain. In addition, IRT analyses were performed for the total set of items to explore IRT-based cross-calibration between the total score on the full-length WRFQ and the short versions (please see below). The IRT parameters are defined relative to the mean and standard deviation of the combined general and shift workers sample (set to have mean 0 and standard deviation 1). The discrimination parameter reflects the item's ability to distinguish between work role functioning levels, with higher scores indicating better discrimination ability. The assumption of a rank order of item response categories was tested in initial analyses using a nominal categories model [26]. If the nominal categories model supported a rank order of response categories, a graded

Table 3. Summary of item properties

WRFQ item	Missing[a]	Ceiling problems[b]	CFA[c]	IRT Disc[d]	IRT fit[e]	DIF[f]	Translatability[g]	10-item version	5-item version
Work scheduling demands									
1. Get going easily	0	1	0	16	25	0	1		
2. Start as soon as you arrived	**0**	**1**	**0**	**5**	**12**	**0**	**0**	**x**	**x**
3. Work without stopping	0	1	0	24	23	0	0		
4. Stick to a routine	0	1	0	25	15	0	0		
Output demands									
5. Work fast enough	*0*	*1*	*0*	*11*	*22*	*3*	*0*	*x*	
6. Finish work on time	0	0	0	15	27	0	0		
7. Work without mistakes	**0**	**1**	**0**	**9**	**26**	**0**	**0**	**x**	**x**
8. Satisfy people who judge	0	1	0	10	11	0	0		
9. Feel accomplishment	0	0	1	27	16	0	1		
10. Done what you are capable of	0	1	1	22	17	0	0		
Physical demands									
11. Lift, carry, or move	1	1	1	17	3	8	0		
12. Stay in one position	*0*	*1*	*0*	*12*	*4*	*0*	*0*	*x*	
13. Repeat the same motions	**1**	**1**	**0**	**3**	**2**	**0**	**0**	**x**	**x**
14. Bend, twist, or reach	*1*	*1*	*0*	*2*	*9*	*2*	*0*	*x*	
15. Use hand-held tools	0	1	1	20	6	5	0		
Mental and social demands									
16. Keep your mind on your work	0	0	0	7	8	0	0		
17. Do work carefully	*0*	*1*	*0*	*8*	*19*	*0*	*0*	*x*	
18. Concentrate on your work	**0**	**0**	**1**	**1**	**5**	**0**	**0**	**x**	**x**
19. Not losing your train of thought	0	0	1	4	18	0	1		
20. Easily read or use your eyes	0	0	0	18	7	0	0		
21. Speak with people	0	1	0	23	24	0	0		
22. Control your temper	0	1	0	26	20	3	1		
Flexibility demands									
23. Set priorities	*1*	*1*	*0*	*13*	*14*	*0*	*0*	*x*	
24. Handle changes	1	1	0	19	13	0	0		
25. Process incoming information	1	0	1	21	10	3	0		
26. Perform multiple tasks	**1**	**0**	**0**	**6**	**1**	**0**	**0**	**x**	**x**
27. Be proactive, show initiative	1	1	0	14	21	0	0		

Both the 5 and 10 item versions are indicated in bold.
10 item version is indicated in italics.
[a] Items with lowest proportion of missing or "not relevant" responses (0 = good).
[b] Items with lowest proportion of respondents in one category (0 = good).
[c] Identified for removal during CFA's (0 = good).
[d] Highest IRT discrimination parameter (rank, low = good).
[e] Lowest ratio of G^2/DF (rank, low = good).
[f] % R^2 difference (rank, low = good).
[g] Multiple issues with translations to other languages (low = good).

response model was fitted for the items [27]. For the item pairs 9/10 (feel a sense of accomplishment/feel you have done what you are capable of doing) and 18/19 (concentrate on your work/work without losing train of thought) where CFA analyses had shown local dependence, item parameters were estimated in two separate runs, excluding the other item in the pair. The difficulty parameters indicate the thresholds on the scale for picking a higher item response. For example, the difficulty 4 parameter indicates the IRT score threshold above which respondents tend to pick the best response "difficult none of the time" rather than responses indicating difficulties "some/half/most of the time". Item fit was evaluated using the S-G^2 test statistic [28]. Items were ranked, based on the IRT discrimination parameters, with lower rank scores equaling good discrimination and higher rank scores indicating poor discrimination. Items were also ranked based on item fit evaluated as the ratio of S-G^2 and degrees of freedom (df), with lower rank scores equaling a good fit and higher scores equaling a poor fit.

Table 4. Comparison of simple scoring and IRT cross-calibration for WRFQ5 and WRFQ10

		Results stratified according to data set					Results stratified according to self-rated health				
	All	Shift workers	University workers	General population	Occupational and insurance physicians	Cancer patients	Excellent	Very good	Good	Fair	Poor
WRFQ	***n* = 2,275**	***n* = 1,055**	***n* = 284**	***n* = 553**	***n* = 154**	***n* = 229**	***n* = 209**	***n* = 605**	***n* = 1,112**	***n* = 239**	***n* = 16**
WRFQ score (0–100)	84.5	86.7	84.1	84.0	82.8	77.1	89.3	88.5	83.7	75.0	64.5
WRFQ5											
Signed difference											
Simple sum score	−0.2	−0.4	0.2	−0.3	1.1	−0.1	−1.1	0.4	−0.3	−0.3	0.0
IRT cross-calibration	0.6	0.1	0.7	0.5	1.9	1.9	−0.9	0.4	0.6	2.2	3.4
Absolute difference											
Simple sum score	3.9	3.5	3.7	4.1	3.7	5.2	3.2	3.3	4.0	5.4	4.4
IRT cross-calibration	3.7	3.4	3.6	3.9	3.5	4.8	3.0	3.3	3.7	5.2	5.6
Square difference											
Simple sum score	29.4	24.6	25.3	32.1	23.5	51.9	25.4	21.3	30.5	49.4	29.5
IRT cross-calibration	25.7	21.8	24.2	27.1	21.5	43.6	18.4	20.0	25.3	47.2	47.9
WRFQ10											
Signed difference											
Simple sum score	0.5	0.3	1.0	0.4	2.2	0.0	0.2	0.9	0.4	0.4	0.5
IRT cross-calibration	0.2	−0.1	0.7	0.1	1.8	0.1	−0.3	0.4	0.1	0.6	1.1
Absolute difference											
Simple sum score	3.0	2.7	2.4	3.1	3.3	4.0	2.4	2.5	3.1	3.9	5.1
IRT cross-calibration	2.9	2.6	2.5	3.2	3.1	3.8	2.5	2.4	3.0	3.8	5.0
Square difference											
Simple sum score	16.7	14.5	12.3	18.1	17.8	26.9	15.4	12.1	17.0	26.9	48.1
IRT cross-calibration	15.5	13.1	12.5	17.4	16.6	24.3	14.6	11.2	15.7	25.5	42.1

2.6. Selection of items for short versions

All candidate items were evaluated with respect to the item level statistics and additionally with respect to evaluation of translatability. All items received a score regarding issues with previous translations and adaptations to other languages and cultures (0 represents no issues with translatability and 1 represents issue(s) with translatability) [19,29–31]. The best scoring items within each domain were selected for inclusion in the shorter versions. The statistical results could be overruled for items deemed conceptually important.

To preserve content validity and with the aim to obtain comparability of scores for the full-length questionnaire and the short versions, an initial 5-item short version was developed by selecting one item from each of the five domains identified in previous factor analyses: 1) Work scheduling, 2) output demands, 3) physical demands, 4) mental and social demands, and 5) flexibility demands. A 10-item short version was developed by selecting additional items from the five domains to determine whether a psychometric performance contrast exists. One should expect a 10-item version to perform better but the question is how much better.

2.7. Mapping to the total score of the full-length 27-item WRFQ v2.0

To calculate sum scores for the two short versions, the same rules are applied as for the full-length 27-item version

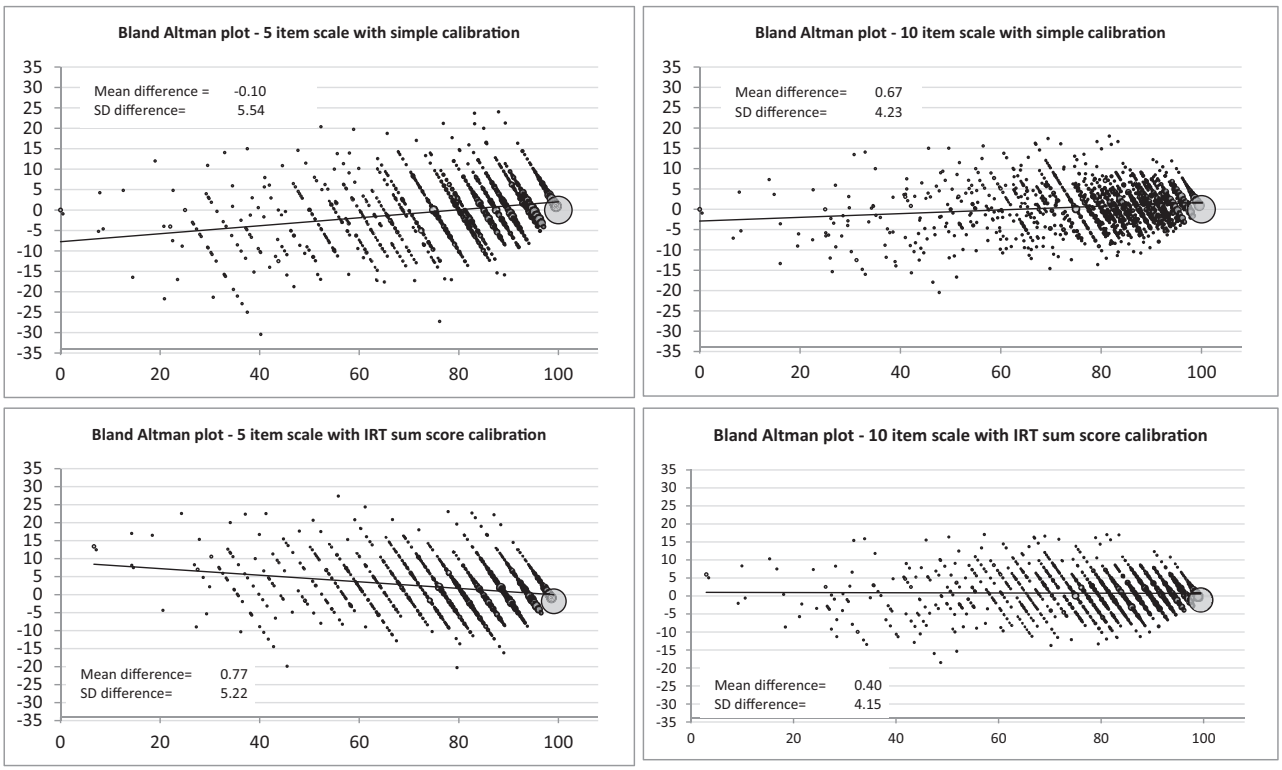

Fig. 1. Bland-Altman plots of agreement between short versions and full WRFQ v2.0.

[4]. A score was calculated if at least 80% of the items were answered. Different scoring methods were explored and compared regarding the abilities of the two short versions to reproduce the total score of the 27-item WRFQ v2.0: 1) simple summative scoring and 2) IRT-based sum score cross-calibration [32]. The latter technique has the advantage of the straightforward linking process built into IRT methodology as well as the utility and practicality of comparing different versions on the summed-score. Similarly to the 27-item version, the short version scores were transformed to scores from 0 (worst work role functioning) to 100 (best work role functioning). Evaluation of agreement between the two shorter versions and the 27-item version was based on the mean signed difference, mean absolute difference, and mean squared difference. These statistics were calculated in the total sample and in subgroups by data set and levels of self-rated health. Furthermore, we evaluated Bland-Altman plots and interclass correlations (ideally >0.7 for scales to be used on group level and >0.9 for scales to be used on individual level [33]). In these analyses, the common mental disorder population sample was excluded, because flexibility demands items were not administered in this sample.

2.8. Scale-level measurement properties

Construct validity was assessed through evaluation of prespecified hypotheses with respect to: 1) job type (manual/nonmanual)—hypothesis: manual job = lower WRFQ scores, 2) self-rated health (excellent-very good-good/fair-poor)—hypothesis: better health = higher WRFQ scores, and 3) number of chronic diseases: (0/1/2/≥3)—hypothesis: more chronic diseases = lower WRFQ scores. Scores were compared using analysis of variance. The performance of the 5- and 10-item versions was compared to the performance of the 27-item WRFQ v2.0, with the hypothesis that the short versions behave similar to the 27-item version.

3. Results

3.1. Item-level analyses

All items and subscales showed ceiling problems and were skewed to the right, especially for the physical demands subscale. For 19 items, more than 50% of respondents selected the best possible score. Missing items were most frequent in the physical demands and flexibility demands subscales. No DIF was found for gender or age, but several items showed significant DIF across populations. Items 5 (work fast enough), 11 (lift, carry, or move objects >5 kilo), 14 (bend, twist, or reach), 15 (use hand-held tools), 22 (control your temper), and 25 (process incoming information) showed an R^2 difference exceeding 2%. DIF across populations was particularly found for items in the physical demands subscale.

In IRT analyses, the rank order assumption was fairly well fulfilled, except for the worst response category "all of the time" which did not discriminate well. For most of

Table 5. Hypotheses testing WRFQ short versions

	Self-rated health				Job type		
Sample	**Good (*N* = 484)**	**Poor (*N* = 58)**	**F**	***P***	**Manual (*N* = 156)**	**Non-M (*N* = 251)**	**F**
General workers							
WRFQ27	**85.2 (14.7)**	**75.5 (21.2)**	**10.9**	**0.002**[b]	84.1 (18.9)	86.8 (12.8)	2.5
WRF10	**85.6 (15.2)**	**75.2 (21.6)**	**12.5**	**0.001**[b]	**82.9 (19.8)**	**87.1 (13.2)**	**5.5**
WRF5	**84.9 (16.4)**	**73.9 (22.8)**	**12.7**	**0.001**[b]	82.7 (20.9)	86.1 (15.0)	3.1
	Good (*N* = 853)	Poor (*N* = 90)	F	*P*	Manual (*N* = 247)	Non-M (*N* = 87)	F
Shift workers							
WRFQ27	**88.1 (12.8)**	**75.1 (75.1)**	**46.3**	**0.000**[b]	88.7 (10.2)	84.7 (16.6)	3.2
WRF10	**88.3 (14.0)**	**76.0 (18.2)**	**38.7**	**0.000**[b]	**89.5 (10.9)**	**83.3 (18.4)**	**8.9**
WRF5	**87.5 (14.3)**	**75.7 (18.1)**	**34.6**	**0.000**[b]	**88.7 (11.7)**	**82.4 (18.8)**	**8.1**
	Good (*N* = 170)	Poor (*N* = 56)	F	*P*	Manual (*N* = 23)	Non-M *N* = 139)	F
Cancer patients							
WRFQ27	**79.2 (16.5)**	**71.1 (20.2)**	**8.3**	**0.004**	**84.1 (13.1)**	**74.7 (18.1)**	**5.1**
WRF10	**78.9 (18.0)**	**72.1 (20.7)**	**5.3**	**0.022**	81.2 (14.7)	75.3 (19.3)	3.9
WRF5	78.5 (19.8)	73.0 (21.1)	3.0	0.086	**84.3 (14.5)**	**74.3 (20.7)**	**3.7**
	Good (*N* = 130)	Poor (*N* = 11)	F	*P*			
Occupational and insurance physicians							
WRFQ27	83.5 (12.8)	76.9 (12.9)	2.5	0.117			
WRF10	85.7 (12.7)	79.5 (9.1)	2.5	0.114			
WRF5	84.8 (14.2)	77.0 (13.7)	3.0	0.085			
	Good (*N* = 229)	Poor (*N* = 29)	F	*P*			
University workers							
WRFQ27	**86.7 (12.7)**	**70.0 (18.1)**	**23.2**	**0.000**[b]			
WRF10	**87.5 (12.5)**	**72.3 (18.0)**	**19.5**	**0.000**[b]			
WRF5	**87.0 (13.6)**	**69.9 (19.6)**	**21.5**	**0.000**[b]			

Significant *P* values <0.05 are in bold.
[a] No chronic condition vs. 1 vs. 2 vs. ≥ 3.
[b] Significant Levine's test for equality of variances (=variances differ significantly).

the subsequent analyses, this lack of discrimination did not pose major problems, but the items regarding work scheduling demands showed very poor model fit. The fit was improved somewhat by collapsing the two worst response categories "all of the time" and "most of the time" for items in this subscale (data available from first author on request).

Table 2 shows the IRT item parameter estimates and item fit statistics from the graded response model. For the items on work scheduling, the two worst categories were collapsed, thus the threshold for the best item category is difficulty 3. Most discrimination parameters were high and some were very high, up to 6.51. For most items, difficulty 4 was around or below 0, reflecting that at least half the respondents tended to choose the best response category on most items. Many items showed poor fit to the IRT model. This was particularly the case for items in the work output demands subscale. Items in the physical demands and flexibility demands subscales generally had acceptable fit, whereas results were mixed for the work scheduling and mental and social demands subscales.

3.2. Selection of items for short versions

Table 3 shows the summary of the information available for item selection and the choices for the 5-item and 10-item short versions. For the 5-item short version, the item with the best overall ranking of item properties was selected within each of the 5 conceptual domains (items 2, 7, 13, 18, 26). The overall principle for the 10-item short version was to select the next best item within each domain. However, practical and conceptual considerations caused some deviations from this principle. Because of the poor fit to the IRT model of the remaining work scheduling items and their relatively low discrimination, no additional work scheduling item was included in the 10-item version. Instead, we included the two next best items concerning physical demands (items 12 and 14) based on conceptual

Table 5 ***(continued)***

Job type	**Chronic conditions**[a]					
P	**0 (*N* = 160)**	**1 (*N* = 72)**	**2 (*N* = 23)**	**≥3 (*N* = 18)**	**F**	***P***
0.115[b]	**89.3 (11.7)**	**83.4 (14.9)**	**82.1 (15.4)**	**76.9 (18.6)**	**7.1**	**0.000**
0.019[b]	**89.7 (11.8)**	**84.5 (17.5)**	**80.0 (16.1)**	**74.7 (19.7)**	**9.5**	**0.000**
0.080[b]	**88.9 (13.4)**	**82.9 (15.2)**	**87.4 (17.9)**	**75.2 (21.6)**	**7.1**	**0.000**
P						
0.078[b]						
0.004[b]						
0.005[b]						
P	0 (*N* = 112)	1 (*N* = 66)	2 (*N* = 32)	≥3 (*N* = 19)	F	*P*
0.026	**80.1 (15.8)**	**76.1 (17.5)**	**76.4 (19.5)**	**65.4 (21.3)**	**3.8**	**0.011**
0.172	**80.6 (16.2)**	**76.3 (18.9)**	**74.6 (22.2)**	**64.9 (21.4)**	**4.2**	**0.007**
0.031	**80.2 (17.4)**	**76.5 (21.3)**	**75.0 (23.2)**	**64.6 (21.5)**	**3.5**	**0.017**

considerations, and the fact that this domain often has the most missing items (see Table 3). In the mental and social demands domain, the best additional items (items 16 and 19) were deemed conceptually too close to the first item chosen. Instead item 17 was chosen.

3.3. Mapping to the score of the 27-item WRFQ v2.0

When comparing the simple summative scores of the short versions to the 27-item total score, mean score differences were close to zero in all data sets tested, both for the total sample as well as the subgroups by data set and levels of self-rated health (Table 4). Bland-Altman plots showed that the simple scoring of the short versions provided lower scores than the 27-item total score for low overall scores (Figure 1). In general, however, the agreement between the 5-item and the 10-item short versions and the 27-item score was acceptable (Figure 1). IRT-based sum score cross-calibration did not lead to noticeable improvement in agreement for the 5-item short version but to slight improvement in agreement for the 10-item short version, in particular for low overall scores. All interclass correlations were >0.93.

3.4. Scale-level measurement properties

Table 5 shows the results of comparisons between several known groups and their WRFQ simple summative scores for both the full-length version and the two short versions. The 10-item short version provided most statistical power in 6 of 8 comparisons that were statistically significant (based on F-value). In these comparisons, the 5-item short version provided similar or better statistical power than the full 27-item version. However, in the cancer patients, the only population with a specific diagnosis, the 5-item short version did not show a significant difference between respondents with poor and good self-rated health, whereas the 10-item short version and the full-length 27-item version did. In the shiftwork population, the full-length questionnaire did not show a statistically significant

difference in work role functioning between job types, while both short versions did. In comparison across all categories, for all questionnaire versions, lower WRFQ values are observed for workers with more chronic conditions compared to workers with less chronic conditions.

4. Discussion

Our study aim was to develop and validate a short version of the WRFQ v2.0 reflecting the psychometric properties of the full-length 27-item version and with the same ability to discriminate between known groups. Using both classical test theory and IRT methods, two short versions with 5 and 10 items were developed. Although items were selected to reflect all five domains of the full-length WRFQ v2.0, we have not pursued subscale scoring because of the brevity of the short versions. The 10-item short version showed better concordance with the full-length WRFQ and better comparability in known groups comparisons (validity) compared to the 5-item short version but at the cost of 5 additional items.

The various methods were able to identify potential items for removal because of their measurement properties. Several items were identified by multiple methods, indicating the robustness of the findings. However, the final decision for item selection was based on both psychometric results and conceptual considerations. These considerations were mainly based on a good representation of the subscale construct and item translatability. For example, item 7 (work without mistakes) was chosen over other items to be included in the two short versions because this item was considered to better reflect the output demands compared to the other items in the subscale, even though this item scored poor on IRT fit.

The two short versions showed acceptable agreement with the total score of the full-length 27-item version. It should be noted, however, that previous research recommends the use of the subscale scores rather than the total score because of the different second-order loadings in the various samples [20]. The short versions are scored with a single summative score, not with subscales. However, this does not imply that the reflective nature of the construct is no longer assumed. The short versions might be good screening instruments, but to get a full understanding and ability to compare between different groups and samples, we recommend using the full-length questionnaire with subscale scores. For use as a screener, more research is needed to develop cutoff scores for both the full-length and the short versions, not only based on statistical considerations but also incorporating clinical and workplace meaningful differences between groups and over time. Our IRT analyses showed that the two worst response categories (“difficult all of the time” and “difficult most of the time”) could be collapsed without any reduction in item performance for the work scheduling demands scale in the combined populations of shift workers and the general population. Further research is needed in other populations, especially clinical populations, to further explore the possibilities for adapting the response categories. Across items, the highest threshold parameter was close to zero, reflecting that approximately half of the respondents chose the best item response category. This ceiling problem is well known in scales for work role function [20,34], reflecting that approximately half of a normal working population do not assess that their health poses any limitations in their work role function.

Even though multiple short versions of lost productivity at work questionnaires exist [6], for example, the 6-, 8-, and 16-item versions of the work limitations questionnaire [35] and the 6- and 13-item versions of the Stanford presenteeism scale [7,36], the development is often not well documented in the literature. In addition, the measurement properties of the different versions in comparison to the full-length versions are often not well studied. With the present study, we provided the first study examining the conversion of a work productivity questionnaire to a shorter version and its measurement properties. Study strengths include the use of multiple working populations and populations with clinical conditions, the rank order of the response categories showing that DIF and IRT analyses are meaningful, and the interplay of measurement properties with conceptual clarity showing strong discriminant validity. Study weaknesses include the inclusion of only two clinical populations, working cancer patients, and workers with common mental disorders, with the common mental disorder population not including the flexibility demands items and therefore left out in several analyses. The limited clinical samples in our study may have implications for the transferability of the results to other clinical populations. Further research is needed to study the behavior of the two short versions in clinical samples. In addition, we found poor fit in the IRT analyses and skewed responses, as are often seen in healthy working populations. The rather high work functioning scores in the shift work population might be explained by a rather healthy population, limited variability between the included shift schedules or the healthy worker effect [21].

In sum, two short versions with 5 and 10 items were identified that are able to reproduce the measurement properties of the full-length 27-item version. The 10-item version performs slightly better in the IRT sum score calibration approach compared to the simple scoring approach (at least concerning agreement with the total score). However, based on the comparison of simple scoring and IRT cross-calibration for both short versions, the simple summative score is recommended, especially given the increased complexity in scoring using the IRT sum score. Both the 5-item and the 10-item versions can be used to measure work role functioning in working samples with mixed clinical conditions and job types. The choice between the 5-, 10-, or 27-item versions depends on the

intended use of the instrument and is a compromise between length and measurement properties.

Acknowledgments

The authors acknowledge Iris Arends, Heleen Dorland, Peter Flach, Hardy van de Ven, and Jac van der Klink for providing their data for the conduct of this study.

References

[1] Tang K, Pitts S, Solway S, Beaton D. Comparison of the psychometric properties of four at-work disability measures in workers with shoulder or elbow disorders. J Occup Rehabil 2009;19:142−54.
[2] Nieuwenhuijsen K, Franche RL, van Dijk FJ. Work functioning measurement: tools for occupational mental health research. J Occup Environ Med 2010;52:1076−2752.
[3] 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.
[4] Abma FI, van der Klink JJ, Terwee CB, Amick Iii BC, Bültmann U. Evaluation of the measurement properties of self-reported health-related work-functioning instruments among workers with common mental disorders. Scand J Work Environ Health 2012;38:5−18.
[5] 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 2014;14:115.
[6] Mateen BA, Doogan C, Hayward K, Hourihan S, Hurford J, Playford ED. Systematic review of health-related work outcome measures and quality criteria-based evaluations of their psychometric properties. Arch Phys Med Rehabil 2017;98:534−60.
[7] Koopman C, Pelletier KR, Murray JF, Sharda CE, Berger ML, Turpin RS, et al. Stanford presenteeism scale: health status and employee productivity. J Occup Environ Med 2002;44:14−20.
[8] Lerner D, Amick BC III, Rogers WH, Malspeis S, Bungay K, Cynn D. The work limitations questionnaire. Med Care 2001;39:72−85.
[9] Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, et al. The world health organization health and work performance questionnaire (HPQ). J Occup Environ Med 2003;45:156−74.
[10] Tang K. Estimating productivity costs in health economic evaluations: a review of instruments and psychometric evidence. Pharmacoeconomics 2015;33:31−48.
[11] Arends I, van der Klink JJ, van Rhenen W, de Boer MR, Bultmann U. Prevention of recurrent sickness absence in workers with common mental disorders: results of a cluster-randomised controlled trial. Occup Environ Med 2014;71:21−9.
[12] Dorland HF, Abma FI, Roelen CA, Smink JG, Ranchor AV, Bultmann U. Factors influencing work functioning after cancer diagnosis: a focus group study with cancer survivors and occupational health professionals. Support Care Cancer 2016;24:261−6.
[13] James L. Redefining work as a result of globalisation and the use of new information technologies. OSHA 2000;2:38−40.
[14] Rantanen J. Research challenges arising from changes in worklife. Scand J Work Environ Health 1999;25:473−83.
[15] Abma FI, van der Klink JJ, Bultmann U. The work role functioning questionnaire 2.0 (Dutch version): examination of its reliability, validity and responsiveness in the general working population. J Occup Rehabil 2013;23:135−47.
[16] Amick BC III, Lerner D, Rogers WH, Rooney T, Katz JN. A review of health-related work outcome measures and their uses, and recommended measures. Spine 2000;25:3152−60.
[17] Amick BC III, Gimeno D. Measuring work outcomes with a focus on health-related work productivity loss. In: Wittink H, Carr D, editors. Pain management: evidence, outcomes, and quality of life: a sourcebook. Amsterdam: Elsevier; 2008:329−43.
[18] Amick BC III, Habeck RV, Ossmann J, Fossel AH, Keller R, Katz JN. Predictors of successful work role functioning after carpal tunnel release surgery. J Occup Environ Med 2004;46:490−500.
[19] Abma FI, Amick BC III, Brouwer S, van der Klink JJL, Bültmann U. The cross-cultural adaptation of the work role functioning questionnaire to Dutch. Work 2012;43:203−10.
[20] Abma FI, Bültmann U, Amick BC III, Arends I, Dorland HF, Flach PA, et al. The work role functioning questionnaire v2.0 showed consistent factor structure across six working samples. J Occup Rehabil 2018;28:465−74.
[21] van de Ven HA, Brouwer S, Koolhaas W, Goudswaard A, de Looze MP, Kecklund G, et al. Associations between shift schedule characteristics with sleep, need for recovery, health and performance measures for regular (semi-)continuous 3-shift systems. Appl Ergon 2016;56:203−12.
[22] Dorland HF, Abma FI, Roelen CAM, Stewart RE, Amick BC, Ranchor AV, et al. Work functioning trajectories in cancer patients: results from the longitudinal work life after cancer (WOLICA) study. Int J Cancer 2017;141:1751−62.
[23] Flach PA. Sick leave management beyond return to work. Groningen: University of Groningen; 2014.
[24] Ware J Jr, Kosinski M, Keller SD. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996;34:220−33.
[25] Zumbo BD. A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.
[26] Bock R. The nominal categories model. In: van der Linden W, Hambleton R, editors. Handbook of modern item response theory. Berlin: Springer; 1997:3−50.
[27] Samejima F. Graded response model. In: van der Linden W, Hambleton R, editors. Handbook of modern item response theory. Berlin: Springer; 1997:85−100.
[28] Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Appl Psychol Meas 2000;24:50−64.
[29] Ramada JM, Serra C, Amick BC 3rd, Castano JR, Delclos GL. Cross-cultural adaptation of the work role functioning questionnaire to Spanish spoken in Spain. J Occup Rehabil 2013;23:566−75.
[30] Gallasch CH, Alexandre NMC, Amick B. Cross-cultural adaptation, reliability, and validity of the work role functioning questionnaire to Brazilian Portuguese. J Occup Rehabil 2007;17:701−11.
[31] Durand MJ, Vachon B, Hong QN, Imbeau D, Amick BC III, Loisel P. The cross-cultural adaptation of the work role functioning questionnaire in Canadian French. Int J Rehabil Res 2004;27:261−8.
[32] Orlando M, Sherbourne CD, Thissen D. Summed-score linking using item response theory: application to depression measurement. Psychol Assess 2000;12:354−9.
[33] Nunnally JC, Bernstein IH. Psychometric theory. New York: McGraw-Hill; 1994.
[34] Maruish ME, editor. User's manual for the SF-36v2 Health Survey. 3rd ed. Lincoln, RI: QualityMetric Incorporated; 2011.
[35] 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.
[36] Turpin RS, Ozminkowski RJ, Sharda CE, Collins JJ, Berger ML, Billotti GM, et al. Reliability and validity of the Stanford presenteeism scale. J Occup Environ Med 2004;46:1123−33.