5. PAPER # 4

Responsiveness of the Work Role Functioning Questionnaire (Spanish version). Journal of Occupational and Environmental Medicine. [In Press 2013].

TITLE:

Responsiveness of the Work Role Functioning Questionnaire (Spanish version) in a general working population.

Running title: Work Role Functioning Questionnaire Responsiveness

AUTHORS:

José M Ramada Rodilla, MD, MSc[1,2,3]
George L Delclós Clanchet, MD, MPH, PhD[1,3,4]
Benjamin C Amick, PhD[4,5]
Femke I Abma, PhD[6]
Juan R Castaño Asins, MD[7]
Gemma Pidemunt Moli, MD, PhD[8]
Ute Bültmann, PhD[6]
Consol Serra Pujadas, MD, PhD[1,2,3]

AFFILIATIONS:

[1] Centro de Investigación en Salud Laboral (CiSAL), Universidad Pompeu Fabra, Barcelona, España.
[2] Servicio de Salud Laboral, Parc de Salut MAR, Barcelona, España.
[3] CIBER de Epidemiología y Salud Pública (CIBERESP).
[4] Southwest Center for Occupational and Environmental Health, The University of Texas School of Public Health. Houston, Texas, USA.
[5] Institute for work & Health. 80 University Avenue, Toronto, Ontario, Canada.
[6] Department of Health Sciences, Work & Health, University Medical Center Groningen, University of Groningen. Groningen, The Netherlands.
[7] Psychiatry Service. Parc de Salut MAR. Hospital del Mar. Barcelona, Spain.
[8] Orthopedic Surgery and Traumatology Service. Parc de Salut MAR. Hospital del Mar. Barcelona, Spain.

CORRESPONDING AUTHOR:

José Mª Ramada Rodilla

Occupational Health Service. Parc de Salut Mar.

Passeig Marítim, 25-29. 08003-Barcelona. Spain

Email address: jramada@parcdesalutmar.cat

Telephone number: +34 932483066;

Fax number: +34 933160410

CONFLICTS OF INTEREST AND SOURCE OF FUNDING:

This project was supported by a grant from the "Fondo de Investigaciones Sanitarias (FIS: PI12/02556), Instituto de Salud Carlos III, Subdirección General de Evaluación y Fomento de la Investigación, Ministerio de Ciencia e Innovación, Spanish Government".

Conflicts of interest: none declared.

CLINICAL SIGNIFICANCE

The study provides evidence that the WRFQ-SpV is an appropriate instrument to measure (true) changes in health-related work functioning over time. However, more research is needed to assess the ability of the instrument detecting (true) changes in groups whose health is stable or deteriorates in individualized health-related work functioning surveillance.

ABSTRACT

Objective: To examine the responsiveness of the Work Role Functioning Questionnaire Spanish-Version (WRFQ-SpV) so that it could be used in evaluative studies.

Methods: A longitudinal survey was performed. Combinations of distribution and anchor-based approaches were used. Five hypotheses were tested examining validity of change-scores. The consensus-based standards for the selection of health status measurement instruments (COSMIN) guided the study design.

Results: One hundred and two participants (mean age=47.3; SD=10.3) completed the WRFQ-SpV twice, within a mean interval of 3.7 months (SD=1.8). Four hypotheses were confirmed and one was rejected. It was verified that the WRFQ-SpV was able to detect (true) changes over time.

Conclusion: Suggestive evidence about the possible use of the WRFQ-SpV with evaluative purposes was provided. More research is needed to examine the instrument responsiveness for groups whose health is stable or deteriorates.

Key terms: responsiveness; measurement instrument; work-functioning instrument.

INTRODUCTION

Increasing life expectancy and delayed retirement age are creating an older active workforce. Many of the older workers likely have health issues due to chronic diseases that require some form of adaptation while working [1,2]. A paradigm shift is needed in occupational health care settings to sustain a productive labour force throughout a workers career. Helping workers to remain healthy at work should be an important target for occupational health research and practice [3].

Work functioning (WF) is determined by the joint influence of work and health and should be viewed as a continuum rather than a dichotomy [4]. Health is a dynamic concept that changes over time. Regarding sickness absence, effective workplace and/or individual interventions could contribute to an early and sustainable return to work.

Both, the interventions and the persons should be evaluated and monitored in this process. Traditional work outcome measures, such as “present versus absent” are no longer sufficient [5]. Validated instruments are required to evaluate the impact of health on WF and quality instruments are needed to examine “functioning at work” and to perform health-related WF surveillance.

A number of self-reported workplace productivity measurement instruments and work-role specific functioning questionnaires have been developed and compared [6,7,8,9]. Recently, the Work Role Functioning Questionnaire (WRFQ) [4], designed to measure health-related WF and freely available in the literature for professionals and researchers, has been successfully translated, adapted and validated to be used in different languages [10-14].

Recent reviews have found that health-related work outcome measures and health-related WF instruments need better validation to make them more meaningful for researchers, practitioners and patients [15,16]. If an instrument is only used to discriminate between patients at one point in time, responsiveness of

the tool is not usually an issue. But for evaluative purposes, when it is intended to measure change in longitudinal studies, this property is very relevant [17].

Responsiveness is the ability of an instrument to detect true change over time in the construct measured. It refers to the validity of a change score, which is the difference between two scores estimated on the basis of at least two measurements [17,18]. Responsiveness of the WRFQ has only been examined in the Dutch version. The Dutch researchers recommended further responsiveness assessments in samples expected to experience changes over time in either work conditions or health status [14].

The ultimate goal of many interventions in occupational health practice and research is to improve WF, and assessing whether a worker's WF status has changed over time is often one of the most important measurement purposes. Therefore, the objective of this study was to examine the responsiveness of the Spanish version of the WRFQ (WRFQ-SpV) so that it could be used in evaluative studies and/or as a monitoring or surveillance instrument.

METHODS

Procedures and sample characteristics

A longitudinal survey was conducted to examine the responsiveness of the WRFQ-SpV. The consensus-based standards on terminology and recommendations to assess the methodological quality of studies on measurement properties of health status measurements instruments (COSMIN) guided the design of the study [18-20].

All participants were recruited, before starting medical treatment, through the outpatient services of psychiatry, physical medicine and rehabilitation, orthopedic surgery and traumatology at a large public hospital system in Barcelona (Spain). Inclusion criteria were: 1) active workers of both sexes, age 18 years and over, 2)

attending his/her first hospital specialist visit, 3) working at least 10 hours per week in the past four weeks, and 4) able to read and understand Spanish. Participants were excluded if they had plans to stop working in the next six months.

Participants were invited to complete the WRFQ-SpV twice on paper (before and after the treatment). At the time of first completion, they provided information on age, gender, level of education (primary, secondary, higher), job type (manual, non-manual, mixed), working hours and health condition (none, musculoskeletal, mental health, others).

One to six months after finalizing treatment at the hospital all participants were invited to answer the WRFQ-SpV again. At that time, a single global perceived effect question (GPE-Q) was added asking respondents to rate their change in WF compared to their pre-treatment baseline, with response options for deterioration from -6 (much worse) to -1 (slightly worse), 0 for no change, and rating improvement from +1 (slightly better) to +6 (much better).

Measure

The WRFQ is a self-administered questionnaire that measures perceived difficulties to perform the job due to health problems [4]. Instructions to use the instrument have been described elsewhere [12]. The WRFQ-SpV consists of 27 items, grouped into 5 subscales reflecting different work demands (work scheduling, output, physical, mental and social). Each item is scored on a Likert five-point scale, anchored to percentages of working time with difficulty handling certain parts of the job. Response options 0, 2 and 4 are anchored to 0%, 50% and 100% respectively. Response option ’does not apply to my job’ is transformed into a missing value. Total scales and/or subscales containing more than 20% missing values are considered missing.

Statistical analysis

The WRFQ-SpV median scores, ranges, mean change scores and standard deviations of change (SD change) were determined. The standard error of measurement (SEM) and Cronbach alpha for the overall scale and each subscale were calculated to evaluate the reliability of the questionnaire. Floor and ceiling effects were explored. These effects occur when more than 15% of responses to a certain item cluster at the top or the bottom of the scale [21].

According to de Vet et al. [17] and the COSMIN panel [18-20] responsiveness is an aspect of validity, and its assessment should emphasize evaluating the validity of change scores. Therefore, analogous to validity, testing hypotheses formulated a priori (before data collection and analysis), concerning the expected correlations and/or expected relationships in different groups measured with the instrument and the GPE-Q, is considered an appropriate method.

A distribution approach was used to estimate the Minimal Important Change (MIC) and to evaluate the hypotheses based on SEM and Effect Size (ES). For those hypotheses addressing correlations and expected change scores in different subgroups of participants an anchor-based approach with the GPE-Q was used.

The following five hypotheses were formulated:

Hypothesis 1: Changes in WF were expected in participants as a result of the treatment, so it was hypothesized that changes in WF over the period of treatment, assessed by the GPE-Q correlate moderate to strongly with the change scores assessed by the WRFQ-SpV. Correlations were assessed using Pearson's correlation coefficient (r), interpreting $r \leq 0.4$ ='weak'; $0.4 \leq r \leq 0.7$ ='moderate' and $r > 0.7$= 'strong' [22].

Hypothesis 2: Treatment was expected to result in improvement, not deterioration, although both were possible. Hence, it was hypothesized a substantially greater effect size (ES) for improvement of WF than for deterioration.

To examine this hypothesis, Cohen's d effect size (ES = difference of means divided by the pooled SD) and the standardized response mean (SRM = mean change divided by SD of change) were calculated as an estimate of magnitude of change over time. Cohen's d ES thresholds were used for interpretation of Cohen's d effect size. Respondents were categorized as deteriorating group (-6 to -1) or improving group (+1 to +6), because statistical methods underlying the SRM assume that all participants change in the same direction [23].

Small ES values ($0.20 \leq ES \leq 0.50$) were hypothesized for participants reporting deterioration in WF, and large ES values ($ES \geq 0.80$) for those reporting improvement.

Hypothesis 3: Positive changes in WF reported on the GPE-Q, predict positive changes in the scores of the WRFQ-SpV. Therefore, it is expected that participants reporting improvement in WF on the GPE-Q, show an "important change" for the overall scale and each subscale of the WRFQ-SpV.

"Important change" was defined as a mean change score larger than both the MIC and the SEM, because responsiveness was being examined for evaluative purposes [21]. Statistical significance of the differences between the WRFQ-SpV scores found at baseline and at follow-up was assessed by means of paired t tests.

Since the WRFQ is a Patient Reported Outcome (PRO), the MIC was considered from the perspective of the patient, and was therefore defined using the GPE-Q as an anchor, as the smallest change of the WRFQ-SpV score which participants perceive as minimally important [17]. The MIC value for improvement was set at

the mean change score of participants reporting an improvement from +1 (slight improvement) to +2 (some improvement).

Hypothesis 4: Participants, who received treatment for physical issues, reporting improvement on the GPE-Q, show an "important change" for the WRFQ-SpV subscale of physical demands (defining "important change" as a change score larger than both, the MIC and the SEM). Statistical significance of change was assessed by means of paired t tests.

Hypothesis 5: Participants, who received treatment for mental health issues, reporting improvement on the GPE-Q, show an "important change" for the WRFQ-SpV subscale of mental demands ("important change" defined like in previous hypothesis). Statistical significance of change was assessed by means of paired t tests.

The contents of the study and the informed consent form were reviewed and approved by the Clinical Research Ethical Committee of the Parc de Salut Mar (Barcelona, Spain) and respect all the principles of the Declaration of Helsinki and Spanish legal regulations on protection of personal data. All analyses were performed with SPSS (Version 15.0. Chicago, IL; 2006).

RESULTS

Sample characteristics:

Table 1 shows the study sample characteristics. A total of 102 participants with a mean age of 47.3 years (SD=10.3) completed the WRFQ-SpV and were included in the analyses. All participants answered the questionnaire twice within a mean interval of 3.7 months (SD=1.8). All were active employees working an average of 39.6 hours per week (SD=7.2), with various levels of education, job types and health problems. Women and participants with high educational level were over

represented in this sample, compared to the general working Spanish population [24].

Median scores, ranges, mean change scores, SD change, SEM and Cronbach alpha for the overall scale and each subscale are presented in table 2.

Floor and ceiling effects:

The overall scale did not show floor or ceiling effects. Subscales did not show floor effects, but ceiling effects were found for the subscales of work scheduling (15%), mental (27%) and social demands (24%) [21] (table 2).

Responsiveness by means of hypothesis testing:

Hypothesis 1: Correlations between the GPE-Q scores and the overall change scores on the WRFQ-SpV for subgroups of participants reporting improvement or deterioration were r=0.5 (p=0.001) and r=0.6 (p=0.002), respectively. Correlations for all subscales and the overall scale in each subgroup were also above 0.4, except for the subscale of social demands in the subgroup of participants improving (table 3).

Hypothesis 2: ES values and SRM are presented in table 4. For the evaluation of change in WF, 34 participants reported deterioration by means of the GPE-Q (Mean= -3.53; SD=1.64), obtaining a mean change score in the WRFQ-SpV of -8.45 (SD=12.67). For those reporting deterioration, Cohen's d ES and SEM were -0.34 and -0.67, respectively. A total of 49 participants reported improvement (GPE-Q Mean=+3.73; SD=1.54) and an average increase in WRFQ-SpV of +22.35 (SD=20.83). For those reporting improvement, Cohen's d ES and SRM were 1.09 and 1.07 respectively, confirming hypothesis 2.

Table 1. Sample characteristics (n=102).

		Total n=102	Men n=45 (44%)	Women n=57 (56%)
Age in years, mean (SD)[a]		47.3 (10.3)	45.9 (9.4)	48.5 (10.8)
Education level, N (%)	Low	28 (27.5)	12 (26.7)	16 (28.1)
	Middle	38 (37.3)	20 (44.4)	18 (31.6)
	High	36 (35.2)	13 (28.9)	23 (40.3)
Working hours/week, mean (SD)[a]		39.6 (7.2)	41.9 (7.4)	37.7 (6.4)
Job type, N (%)	Manual	36 (35.3)	14 (31.1)	22 (38.6)
	Non-manual	27 (26.5)	10 (22.2)	17 (29.8)
	Mixed	39 (38.2)	21 (46.7)	18 (31.6)
Reported health issue type, N (%)	Physical	49 (48.0)	15 (33.3)	34 (59.7)
	Mental health	33 (32.4)	20 (44.4)	13 (22.8)
	Others	20 (19.6)	10 (22.3)	10 (17.5)
Disease duration in months:	Mean, (SD)	22.4 (36.1)	15.2 (16.8)	28.0 (45.3)
	Median, Range	12.0 1-300	12.0 1-60	12.00 1-300

(a): Standard deviation

Table 2. Spanish version of the Work Role Functioning Questionnaire: scores, floor and ceiling effects and Cronbach alpha at baseline. Standard error of measurement, (n=102).

	Valid n[a] (missing/not applicable)	Baseline[b] median scores	Baseline scores range	Baseline n at floor (0%)	Baseline n at ceiling (100%)	Mean change scores / (SD change)[c]	Standard error measurement (SEM)	Cronbach alpha at baseline
Work scheduling demands	98 (4)	80.6	0.00-100	1 (1.0)	15 (15.3)	5.8 (28.0)	19.8	0.93
Output demands	98 (4)	75.0	0.00-100	1 (1.0)	9 (9.2)	9.5 (24.8)	17.6	0.92
Physical demands	74 (28)	70.0	4.17-100	0 (0.0)	10 (13.5)	7.9 (26.3)	18.6	0.95
Mental demands	101 (1)	83.3	0.00-100	3 (3.0)	27 (26.7)	10.2 (24.3)	17.2	0.94
Social demands	88 (14)	83.3	0.00-100	2 (2.7)	21 (23.9)	10.0 (22.9)	11.5	0.93
Total scale	98 (4)	77.9	6.48-100	0 (0.0)	3 (3.1)	8.6 (21.7)	15.0	0.94

(a) Subscales with more than 20% of items scoring "does not apply to my job" or missing values were excluded.

(b) Each subscale is scored from 0-100. Higher scores mean better work functioning: difficulties all the time 0/100; difficulties no of the time 100/100.

(c) SD change: standard deviation of change.

Table 3. Bilateral correlations between the scores of the Global Perceived Effect Question scores and the WRFQ-SpV[a] change scores.

Participants declaring Improvement (n=49)	**Mean Change score WRFQ-SpV[a] (SD)**	**Pearson's (r)[b]**	**Bilateral significance**
Work Scheduling Demands	22.9 (25.6)	0.6	< 0.001
Output Demands	25.6 (23.4)	0.4	0.001
Physical Demands	23.3 (27.0)	0.4	0.165
Mental Demands	23.1 (26.6)	0.4	0.017
Social Demands	20.9 (27.3)	0.3	0.132
Overall Scale	22.3 (20.8)	0.5	0.001
Participants declaring Deterioration (n=31)	**Mean Change score WRFQ-SpV[a]**	**Pearson's (r)[b]**	**Bilateral significance**
Work Scheduling Demands	-24.0 (17.7)	0.6	0.006
Output Demands	-14.3 (10.5)	0.4	0.028
Physical Demands	-17.7 (19.4)	0.5	0.028
Mental Demands	-12.1 (10.6)	0.6	0.047
Social Demands	-16.7 (12.3)	0.4	0.484
Overall Scale	-10.5 (10.8)	0.6	0.002

(a) WRFQ-SpV: Work Role Functioning Questionnaire (Spanish version).

(b) Pearson's correlation coefficient.

Table 4. Assessment of responsiveness by means of the Standardized Response Mean, using Cohen's d Effect Size thresholds for interpretation, (n=102).

	Participants reporting deterioration in WF[a] (N=34)	**Participants reporting improvement in WF[a] (N=49)**
Mean score of the GPE-Q[b] (SD)	-3.53 (1.64)	3.73 (1.54)
WRFQ-SpV[c]mean score at baseline (SD)	71.08 (24.32)	66.09 (25.83)
WRFQ-SpVmean score at follow-up(SD)	62.63 (25.54)	88.44 (11.43)
Mean change (SD); p value[d]	-8.45 (12.60); p=0.001	22.35 (20.83); p<0.001
Standardized response mean (SRM)	-0.67	1.07
Cohen's d effect size (with pooled SD change)	-0.34	1.12

(a) WF: Work functioning (by means of the GPE-Q);
(b) GPE-Q: Global Perceived Effect Question.
(c) Spanish version of the Work Role Functioning Questionnaire.
(d) Paired t test.

Hypothesis 3: Participants reporting improvement in WF on the GPE-Q (n=49), showed mean change scores in all subscales and the overall scale above the MIC values for improvement and the SEM. All change scores in this group were statistically significant ($p < 0.001$) (table 5).

Hypothesis 4: Participants who received treatment for physical issues, reporting improvement on the GPE-Q (n=18), showed a significant change score in the subscale of physical demands (mean change score=19.2; SD=17.0; $p<0.001$), that was above the MIC value for improvement (11.3) and the SEM (12.0), confirming hypothesis 4 (table 5).

Hypothesis 5: Participants who received treatment for mental health issues, reporting improvement in the GPE-Q (n=24), showed a significant change score (mean change score=18.1; SD=28.9; $p=0.006$) that was above the MIC value for improvement (7.4) but below the SEM (20.5), rejecting hypothesis 5 (table 5).

Table 5. Scores on the Spanish version of the Work ole Functioning Questionnaire of participants under treatment, declaring improvement in the GPE-Q, by type of health issue (mental or physical).

	All participants, improvement (N=49)					
	WRFQ-SpV[(a)] Mean (SD)	WRFQ-SpV[(b)] Mean (SD)	Mean change score (SD)	**p value**	**MIC[(c)] improvement**	**SEM[(d)]**
Work scheduling demands	63.4 (30.2)	85.6 (17.3)	24.8 (27.0)*	<0.001	7.3	19.1
Output demands	61.9 (30.8)	86.8 (14.2)	23.9 (31.0)*	<0.001	11.1	21.9
Physical demands	70.1 (27.8)	87.7 (12.4)	20.4 (26.7)*	<0.001	11.3	18.9
Mental demands	68.7 (30.9)	91.8 (10.4)	23.1 (26.6)*	<0.001	7.4	18.8
Social demands	74.6 (26.7)	92.2 (10.8)	20.4 (28.8)*	<0.001	11.1	20.4
Total score	66.1 (25.8)	88.4 (11.4)	22.4 (20.8)*	<0.001	9.0	14.7
	Physical health issues, improvement (N=18)					
	WRFQ-SpV[(a)] Mean (SD)	WRFQ-SpV[(b)] Mean (SD)	Mean change score (SD)	**p value**	**MIC[(c)] improvement**	**SEM[(d)]**
Work scheduling demands	67.6 (28.3)	84.0 (20.6)	16.5 (19.0)	0.003	7.3	13.4
Output demands	64.1 (32.5)	85.0 (18.0)	20.9 (19.3)	<0.001	11.1	13.6
Physical demands	64.6 (26.8)	83.8 (13.7)	19.2 (17.0)*	<0.001	11.3	12.0
Mental demands	74.3 (35.4)	92.6 (12.6)	18.3 (25.3)	0.007	7.4	17.9
Social demands	82.1 (26.3)	89.9 (14.3)	7.7 (18.9)	0.150	11.1	13.4
Total score	68.5 (25.6)	86.7 (13.9)	18.3 (16.0)	<0.001	9.0	11.3
	Mental health issues, improvement (N=24)					
	WRFQ-SpV[(a)] Mean (SD)	WRFQ-SpV[(b)] Mean (SD)	Mean change score (SD)	**p value**	**MIC[(c)] improvement**	**SEM[(d)]**
Work scheduling demands	59.6 (30.3)	88.0 (13.6)	28.3 (30.0)	<0.001	7.3	21.2
Output demands	62.2 (25.9)	81.0 (18.4)	18.8 (28.8)	0.006	11.1	20.4
Physical demands	74.1 (28.1)	85.9 (22.3)	11.8 (17.7)	0.014	11.3	12.5
Mental demands	61.8 (29.3)	79.9 (25.4)	18.1 (28.9)*	0.006	7.4	20.5
Social demands	70.3 (25.5)	80.4 (20.7)	10.1 (27.6)	0.092	11.1	19.5
Total score	61.7 (26.1)	89.0 (9.6)	27.3 (24.2)	<0.001	9.0	17.1

(a) Mean scores at baseline; (b) Mean scores at follow-up (after treatment);

(c) Minimal Important Change (MIC); (d) Standard error of measurement

(*) Hypotheses 3 and 4 confirmed; hypothesis 5 rejected.

DISCUSSION

Four out of five hypotheses were confirmed to examine whether the WRFQ-SpV can be used for evaluative purposes. All hypotheses tested in this study were concerned with the expected changes in participants over time, based on the knowledge of the questionnaire structure and conceptual framework. Consequently, confirming such hypotheses supports the validity of the change scores and hence, the responsiveness of the instrument [17,18,20,21].

All correlations between the WRFQ-SpV and the GPE-Q were confirmed as expected in hypothesis 1, with the exception of the social demands subscale in participants reporting improvement, supporting the validity of the change scores and therefore providing insight about the WRFQ-SpV responsiveness.

Assessing responsiveness by means of ES values and SRM, as it has been carried out in this study to test the second hypothesis, has created some controversy in the literature [17,25]. While many authors have used ES as a measure of responsiveness [26-29] and SRM as one of the more valid measures for its estimation [30], others disagree as they consider ES values as a measure of the *magnitude* of the change scores, rather than its *validity* [17,20,25]. In this study, ES values and SRM were used to test a predetermined hypothesis (grounded in the construct of the questionnaire) about the expected magnitude of the ES for different groups of participants, providing additional understanding on the validity of the change scores, and hence, suggesting that the WRFQ-SpV is a responsive questionnaire.

Hypotheses 3 and 4 were confirmed, but hypothesis 5 was rejected because the mean change score of the WRFQ-SpV for the subscale of mental demands was above the MIC but below the SEM. Considering responsiveness as ‘the ability to detect change in general’ would have lead us to confirm hypothesis 5, but detecting ‘any’ change should not be considered responsiveness because ‘any’

change could refer to true change but also to measurement error or other bias [17,31], and the interest in the study was seeking true changes.

It could be debated what is the best approach to estimate the MIC because it depends on the baseline values of the instrument [32],the type of anchor from the patient or the clinician's perspective [33], the definition of ‘minimal important change’ and the direction of change [17,32]. According to de Vet et al. [17], it is reasonable to consider MIC values from the patient’s perspective for self-reported instruments. Consequently, it was considered appropriate to calculate MIC values from this perspective.

Using receiver operating characteristic (ROC) curves and area under ROC curve (AUC) values to evaluate responsiveness would have provided additional strength to the results. AUC values are interpreted as the probability that a measure correctly discriminates between participants who have improved and those who have not. However this method could not be used because no gold standard is available for WF. Many studies have used the GPE-Q as a gold standard, but it is doubtful the reliability and validity of such retrospective measures of change when used as a gold standard [17,31,34].

It could be argued whether confirming four hypotheses is enough to conclude that the WRFQ-SpV is a responsive instrument or not, or that other hypotheses could had been formulated regarding the expected scores of participants reporting deterioration. This study design was not appropriate to examine such hypotheses. As expected, the number of participants who deteriorated following treatment was small for analysis. Several authors agree that it is not possible to formulate standards on the number of hypothesis that need to be tested because testing responsiveness is a continuous process of accumulating evidence [17,21,31] and this study does provide evidence about the responsiveness of the questionnaire, mainly in participants reporting improvement.

The results of this study on responsiveness are consistent with those obtained for the Dutch WRFQ version, which showed moderate responsiveness. The difference between both studies is that in this study, responsiveness was assessed in a larger group of participants for whom a change was expected due to the treatment. For the Dutch version responsiveness was assessed in a relatively stable and healthy population, with no intervention between the first and the second test, and the number of those reporting change was very small [14].

Ceiling effects were identified for the subscales of work scheduling, mental and social demands. These results are consistent with other studies concerning the WRFQ subscales of mental and social demands [10,12-14]. It seems that certain items of the WRFQ have a lack of discriminative ability to differentiate workers at the higher range of WRFQ scores. It could be argued whether these ceiling effects could affect responsiveness, because participants scoring at the upper part of the scale at baseline cannot show further improvement.

To the authors' knowledge, this is the first study to evaluate responsiveness in a population of workers expected to show changes in WF over time.

A significant strength of this longitudinal study is that responsiveness was examined by focusing more on validity than on the magnitude of the change scores. Consistent with the literature [17,18,25,31], a valid instrument for evaluative purposes or to be used as surveillance instrument should provide evidence of its capacity to measure true changes.

Experts do not agree on a single preferred approach for responsiveness evaluation, and instead recommend combining several approaches, including both anchor and distribution based methods [23,35-38]. Hence, both approaches were combined to make the evaluation consistent with these recommendations.

This study also has limitations. It was carried out with a sample size that was initially considered adequate (n=102) [39]. However, stratification was necessary

and some subgroups were of smaller sample sizes. Further longitudinal studies with a larger number of participants and different health issues are needed, especially to examine responsiveness of the instrument in groups who do not experience improvement, or deteriorate.

The intervention was not uniform for all participants, with a wide range of treatments, from non-invasive (e.g., drug treatment, psychological support, occupational therapy or physical rehabilitation) to invasive (e.g., tendon or joint infiltration, minor outpatient or major surgery), which could cause different degrees of change in participants. As other authors have noted [17,31,40], the challenge in the study design was to find a population with expected changes over time. Therefore, for this study purpose, the main point was to secure a sample of participants in such a way that it would be expected for a proportion of participants to either improve or deteriorate, and then explore whether the questionnaire was able to detect the validity of the changes.

Quality validated instruments are needed for occupational health research and practice. Additionally, evidence on WRFQ responsiveness is needed if it is going to be used for evaluative purposes. This study provides suggestive evidence that the WRFQ is an adequate instrument to measure changes in health-related WF over time. However, more evidence is needed on the ability of the instrument to detect changes in groups who do not experience improvement, or deteriorate during individualized health-related WF surveillance.

REFERENCES

1. Ross D. Ageing and work: an overview. Occup Med (Lond). 2010; 60:169-71.

2. Hairault JO, Langot F, Sopraseuth T. Distance to retirement and older workers employment: the case for delaying the retirement age. J Eur Economic Assoc. 2010;8:1034 -76.

3. Macdonald EB, Sanati KA. Occupational health services now and in the future: the need for a paradigm shift. J Occup Environ Med. 2010;52:1273-7.

4. 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.

5. Abma FI. Work functioning: development and evaluation of a measurement tool [PhD thesis]. Groningen, NL: University of Groningen; 2012. [Internet]. [cited 2013 May 28]; [about 160 p.]. Available from: http://irs.ub.rug.nl/ppn/351176438

6. Mattke S, Balakrishnan A, Bergamo G, Newbwrry SJ. A review of methods to measure health-related productivity loss. Am J Manag Care. 2007;17:547-79.

7. Ozminkowski RJ, Goetzel RZ, Chang S, Long S. The application of two health and productivity instruments at a large employer. J Occup Environ Med. 2004;46:635-48.

8. Prasad M, Wahlqvist P, Shikiar R, Shih YC. A review of self-report instruments measuring health-related work productivity: a patient-reported outcomes perspective. Pharmacoeconomics. 2004;22:225-44.

9. Lofland JH, Pizzi L, Frick KD. A review of health-related workplace productivity loss instruments. Pharmacoeconomics. 2004;22:165-84.

10. 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. 2004;27:261-8.

11. Gallasch CH, Alexandre NM, Amick B 3rd. Cross-cultural adaptation, reliability, and validity of the work role functioning questionnaire to Brazilian Portuguese. J Occup Rehabil. 2007;17:701-11.

12. Ramada JM, Serra C, Amick BC III, Castaño JR, Delclos GL. Cross-cultural adaptation of the Work Role Functioning Questionnaire to Spanish spoken in Spain. J Occup Rehabil. 2013;23:566-75.

13. Abma FI, Amick Iii BC, Brouwer S, van der Klink JJ, Bültmann U. The cross-cultural adaptation of the work role functioning questionnaire to Dutch. Work. 2012;43:203-10.

14. Abma FI, van der Klink JJ, Bültmann 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.

15. Williams RM, Schmuck G, Allwood S, et al. Psychometric evaluation of health-related work outcome measures for musculoskeletal disorders: a systematic review. J Occup Rehabil. 2007;17:504–21.

16. Abma FI, van der Klink JJ, Terwee CB, Amick BC 3rd, 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.

17. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: A practical guide. 1st ed. Cambridge, UK: The University Press Cambridge, 2011.

18. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63:737-45.

19. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Med Res Methodol. 2010;10:22.

20. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19:539-49.

21. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60:34-42.

22. Evans JD. Straightforward statistics for the behavioral sciences. Pacific Grove, CA: Brooks/Cole Pub. Co.; 1996.

23. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Med Care. 2000; 38(9 Suppl):II84-90.

24. Observatorio Estatal de Condiciones de Trabajo [Internet]. VI Encuesta Nacional de Condiciones de Trabajo. [cited 2007]; [about 117 screen, pages 20-23]; Available from:

ttp://www.oect.es/Observatorio/Contenidos/InformesPropios/Desarrollados/Ficheros/Informe_VI_ENCT.pdf

25. Angst F. The new COSMIN guidelines confront traditional concepts of responsiveness. BMC Med Res Methodol. 2011 Nov 18;11:152; author reply 152.

26. Hsueh IP, Hsieh CL. Responsiveness of two upper extremity function instruments for stroke in patients receiving rehabilitation. Clin Rehabil. 2002;16:617-24.

27. Lin JH, Hsu MJ, Sheu CF, Wu TS, Lin RT, Chen CH, et al. Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke. Phys Ther. 2009;89:840-50.

28. Hsieh YW, Wu CY, Lin KC, Chang YF, Chen CL, Liu JS. Responsiveness and validity of three outcome measures of motor function after stroke rehabilitation. Stroke. 2009;40:1386-91.

29. Johansson C, Bodin P, Kreuter M. Validity and responsiveness of the spinal cord index of function: an instrument on activity level. Spinal Cord. 2009;47:817-21.

30. Sivan M. Interpreting effect size to estimate responsiveness of outcome measures. Stroke. 2009;40:e709; author reply e710-1.

31. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th ed. New York, USA: Oxford University Press Inc., 2008.

32. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol. 2003;56:395-407.

33. Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE Jr. Determining minimally important changes in generic and disease-specific health-related

quality of life questionnaires in clinical trials of rheumatoid arthritis. Arthritis Rheum. 2000;43:1478-87.

34. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 1997;50:869-79.

35. Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain. 2008;9:105-21.

36. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol. 1997;50:79-93.

37. Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61:102-9.

38. Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol. 1997;50:239-46.

39. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2011;21:651-7.

40. Beaton DE, Tang K, Gignac MA, Lacaille D, Badley EM, Anis AH, et al. Reliability, validity, and responsiveness of five at-work productivity measures in patients with rheumatoid arthritis or osteoarthritis. Arthritis Care Res (Hoboken). 2010;62:28-37.