J Occup Rehabil (2018) 28:365–376
DOI 10.1007/s10926-017-9728-8

Assessing Psycho-social Barriers to Rehabilitation in Injured Workers with Chronic Musculoskeletal Pain: Development and Item Properties of the Yellow Flag Questionnaire (YFQ)

Cornelia Rolli Salathé[1] · Maurizio Alen Trippolini[2] · Livio Claudio Terribilini[1] · Michael Oliveri[2] · Achim Elfering[1,3]

Published online: 8 September 2017

Abstract *Purpose* To develop a multidimensional scale to asses psychosocial beliefs—the Yellow Flag Questionnaire (YFQ)—aimed at guiding interventions for workers with chronic musculoskeletal (MSK) pain. *Methods* Phase 1 consisted of item selection based on literature search, item development and expert consensus rounds. In phase 2, items were reduced with calculating a quality-score per item, using structure equation modeling and confirmatory factor analysis on data from 666 workers. In phase 3, Cronbach's α, and Pearson correlations coefficients were computed to compare YFQ with disability, anxiety, depression and self-efficacy and the YFQ score based on data from 253 injured workers. Regressions of YFQ total score on disability, anxiety, depression and self-efficacy were calculated. *Results* After phase 1, the YFQ included 116 items and 15 domains. Further reductions of items in phase 2 by applying the item quality criteria reduced the total to 48 items. Phase factor analysis with structural equation modeling confirmed 32 items in seven domains: activity, work, emotions, harm & blame, diagnosis beliefs, co-morbidity and control. Cronbach α was 0.91 for the total score, between 0.49 and 0.81 for the 7 distinct scores of each domain, respectively. Correlations between YFQ total score ranged with disability, anxiety, depression and self-efficacy was .58, .66, .73, −.51, respectively. After controlling for age and gender the YFQ total score explained between R2 27% and R2 53% variance of disability, anxiety, depression and self-efficacy. *Conclusions* The YFQ, a multidimensional screening scale is recommended for use to assess psychosocial beliefs of workers with chronic MSK pain. Further evaluation of the measurement properties such as the test–retest reliability, responsiveness and prognostic validity is warranted.

Keywords Screening · Assessment · Risk factors · Occupational rehabilitation · Back pain · Work

Introduction

The importance of psychological and social factors to risk of development of chronic musculoskeletal (MSK) pain is well-established [1–4]. These psychosocial factors represent major barriers to return to work (RTW), which may cause a substantial burden to the individual and to society [5–8] and, along with potential barriers to rehabilitation, are well described by the *clinical flags approach* [2–4]. All factors are divided into the red, orange, yellow, blue and black flag groups in order to characterize the biological, psychological and social factors affecting health and recovery after injury or illness [3, 9, 10]. For example, in people with unspecific low back pain (lbp), it has been suggested that yellow flags are highly relevant in terms of outcomes compared to the rare red flags that represent bio-medical factors such as structural findings [11]. Specifically targeting the psychosocial factors, so called yellow flags, with interventions may provide better outcomes than ignoring these factors [3]. However, the flags, as proposed by the inventors, are usually assessed by a clinician and therefore may not appropriately reflect the patients' own beliefs. Moreover, it is unclear how

✉ Cornelia Rolli Salathé
cornelia.rolli@psy.unibe.ch

[1] Department of Work and Organisational Psychology, Institute of Psychology, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland

[2] Department of Work Rehabilitation, Rehaklinik Bellikon, SUVA Care, Bellikon, Switzerland

[3] National Centre of Competence in Research, Affective Sciences, University of Geneva, CISA, Geneva, Switzerland

the various psychological risk factors influence outcomes [3].

Due to the limited resources allocated to screening for relevant yellow flags, clinicians may be challenged in their choice of appropriate measures. In order to screen patients with acute lbp early on for potential risk for chronicity, concise self-reported questionnaires have been developed, such as STarT Back Tool (SBT; [12, 13]) or the Orebro Musculoskeletal Pain Screening Questionnaire (OMPSQ; [14]). Their use is aimed at the early phase of the disorder and has shown some limitations when used in other cohorts such as healthy or workers with chronic pain [15].

A plethora of distinct self-report scales for measuring cognitive, emotional or behavioral risk factors have been developed in the last decades. Some of these scales claim to measure beliefs relating to catastrophizing, self-efficacy, coping, fear-avoidance, anxiety and many other beliefs [16–23]. Each of these constructs is usually measured with a separate questionnaire. Hence, carrying out a comprehensive assessment of the most relevant psychological risk factors would require multiple questionnaires, which imposes a considerable burden on both patients and healthcare providers [24, 25].

Therefore, we believe there is a need for a concise screening tool that attempts to measures the multidimensionality of psychological factors that are known to influence outcomes of patients with chronic MSK. Consequently, the aim of this study is to develop and evaluate the Yellow Flag Questionnaire (YFQ).

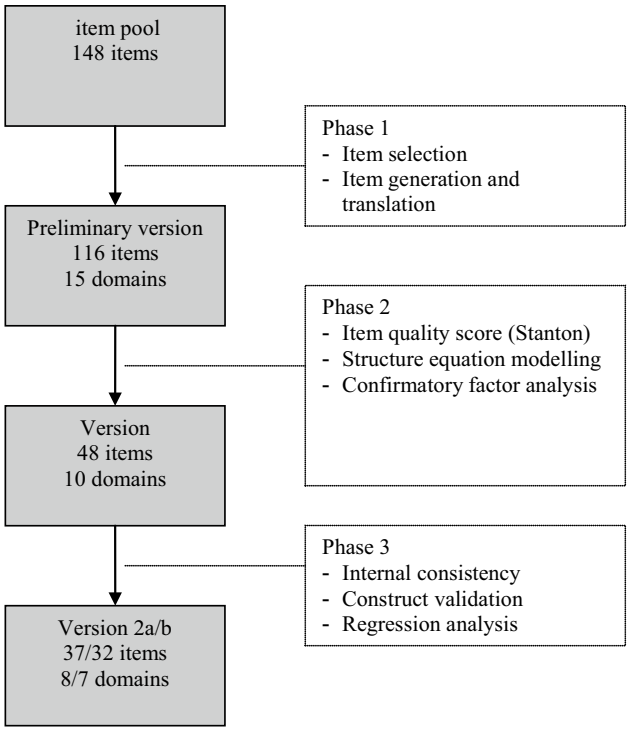

Fig. 1 Development-phases of the various versions of the Yellow Flag Questionnaire

Method

The development of the YFQ followed three phases displayed in Fig. 1. The first phase of the study included the development of a preliminary questionnaire based on a pool of items with corresponding constructs, followed by a second phase comprising the psychometric evaluations. In the third, validation phase, analyses of internal consistency and construct validity were performed.

Phase 1: Development of a Preliminary Version of the YFQ

Generating a Pool of Items

Design An expert consensus method inspired by the Delphi method was performed [26]. A structured process was followed whereby participants revealed and shared their opinion with others. While the assignments from round to round were performed individually, the participants gathered to share opinions from the other participants. Based on the group answers, participants had the opportunity to reconsider their answers until they reached a consensus [26]. The consensus process was coordinated by a member of the research group (MO) until after the second expert round. The data on each round was collected with an Excel spreadsheet that allowed the researchers to keep track of the items, selection criteria and changes during the consensus process.

Included Clinical Experts Clinical experts were gathered based on purposive sampling method [27]. Among the 400 employees of a rehabilitation center in Switzerland, only clinicians working with MSK-patients were invited to participate. All participants were required to have extensive experience in working in an interdisciplinary program focused on the return to work of patients with chronic MSK.

Literature Search and Other Searches The aim of the literature search was to provide an extensive variety of screening factors that would reflect the interdisciplinary character of our gathering and define as many yellow-flag items that were deemed relevant for patients with chronic MSK at risk for non-recovery as possible. By using key words such as “risk or prognostic factors”, “flags”, “screening”, or “chronic pain” to name a few, participants performed a bibliographic literature search in MEDLINE, searched references by hand, and asked other experts in the field. If items were deemed appropriate, they were entered and numbered in the spreadsheet.

First Expert Round The aim of the first round was to select the most appropriate items from the pool gathered in the literature search. For this process, all participants were divided into groups of two experienced clinical experts working independently at first. The selection process was guided by the question: "Which of the following items most relevantly influences rehabilitation and RTW outcomes in patients with MSK?" Afterwards, the two experts discussed their selections, adapted the wording, discussed areas of disagreement and provided a list of items they agreed upon. Lists were collected from all expert-teams and merged into one list. In the following consensus meeting, all participants discussed and finally agreed on the item selection of the preliminary version. Then, a preliminary answer scale which would fit items from various questionnaires was discussed and agreed on.

Second Expert Round With the aim to reduce the number of items to a comprehensive succinct set of items, a second round took place. Items with reversed scoring were marked appropriately. Similar items from distinct questionnaires were blended into one item. If not available in German, items were translated into German by a native German speaker with proficiency in English and proofread by a native English speaker familiar with the German language.

Phase 2: Psychometric Evaluations

Included Participants

Between 2001 and 2008 we collected data from 700 consecutively recruited injured workers with chronic MSK pain undergoing inpatient physical therapy at a rehabilitation center in Switzerland ($M_{age} = 40.7$ years, $SD = 11.2$, 73.3% men). Since the study was embedded in the usual intake procedure, all patients scheduled for a work rehabilitation program were sent the preliminary YFQ via postal mail. They were asked to complete and return it to the rehabilitation center in the enclosed prepaid envelope. On admission to the work rehabilitation program, patients were given the opportunity to discuss the YFQ items with their assigned healthcare provider (doctor or physical therapist), comment on items and amend their responses if appropriate. Patients were only included in the study if they had consented to the use of their data for research purposes. The ethical committee of Aargau canton, Switzerland approved the study.

Assessment of Item Quality

We used the criteria developed by Stanton et al. [28] to determine item quality. Stanton and colleagues describe *internal item qualities, external item qualities*, and *judgmental item qualities*. The first refer to the internal characteristics of items belonging to a single scale, while the external item qualities describe the relations of items or the whole scale to other patient assessments. Judgmental item qualities are assessed subjectively rather than statistically. Because relying exclusively on classical scale reduction methods such as corrected item-total correlations (e.g., [29]) might lead to narrow item content and low validity [30, 31], we used a combination of internal, external, and judgmental criteria to select a pool of high-quality items [28]. Items were excluded if (a) more than 5% of values were missing (i.e., judgmental item qualities); (b) they were subject to a floor or ceiling effect (i.e., internal item qualities); (c) more than 5% of respondents had commented on the item (i.e., judgmental item qualities); (d) the item-subscale total correlation was above 0.4 and the item-other subscales total correlation was below 0.6 (i.e., external item qualities). The analyses were performed by using IBM SPSS (IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.). Final judgments (*yes* vs. *no*) of the eight experts about the clinical importance of the remaining single items were made, again following the Delphi Method [26].

Confirmatory Factor Analyses Testing the Seven Versus the Eight-Domain Model of the YFQ

Regarding the factorial structure of the YFQ, two different models were tested, since the domain pain beliefs might not necessarily constitute an independent domain within the set of yellow flags due to an assessment bias. Most other domains, e.g., beliefs about activity, work and pain control already encompass beliefs about pain [32].

The data were split randomly into two sub-samples ($N_1 = 332$; $N_2 = 332$) to determine which domain structure was the best fit with the data. First, missing values for individual items were categorized as non-ignorable [33] or random missing values. *Non-ignorable missing values* attribute participants' item omission to specific characteristics of interest such as socio-economic status [34]. *Random missing values* do not point out a systematic relationship between participant profile and item omission [34, 35]. Only random missing values were replaced with series means. Second, all items were *parceled* by averaging two or more items scores. In structure equation modelling (SEM), item parcels replace single items as predictors [36]. In non-normally distributed items with a one-dimensional structure, increasing the number of items per parcel can enhance model fit [36] by increasing model parsimony (cf. [37]). However, random methods of combining items into parcels are adequate [37], since random item parceling could reduce measurement error [37, 38]. The unidimensionality of proposed factors was checked using principal component analysis (cf. [37]). Both sub-samples were subjected to confirmatory factor

analysis, integrated into structure equation modelling using AMOS (Version 18).

The following indices of fit for both models were calculated (cf. [39]): model chi-squared (χ^2), χ^2/degrees of freedom, root mean square error of approximation (RMSEA), comparative fit index (CFI) and normed fit index (NFI) [40]. Model chi-squared measures the difference between the fitted covariance matrices and the sample; $p > .25$ suggests good fit. Model χ^2 is sensitive to sample size and tends to be significant for large sample sizes. Chi-squared/df is less sensitive to sample size; $\chi^2/df < 2$ and $\chi^2/df < 3$ indicate good and acceptable fit respectively. RMSEA (acceptable fit: <0.10; good fit: <0.06) provide information about the fit between optimal, yet unknown parameters and the covariance matrix for the population [40]. The CFI and NFI both compare χ^2 for the model to χ^2 for a null/independence model, so values should be as high as possible (good fit: >0.95, acceptable fit: >0.90; [40]). The chi-squared difference test was used to determine whether one model was a better fit to the data than the other: $\chi^2_{7\ \text{factors}}$ (degrees of freedom) $- \chi^2_{8\ \text{factors}}$ (degrees of freedom) $= \Delta\chi^2$ (Δdegrees of freedom) [40].

Phase 3: Validation Analyses

Reliability Analysis

Cronbach's alpha was calculated to estimate internal consistency, a measure of reliability of the YFQ (cf. [40]). Due to the YFQ's multidimensional structure, satisfactory internal consistency was expected at a lower degree compared to single-construct measures, hence Cronbach's $\alpha \geq 0.70$ (cf. [41]).

Construct Validity Analysis

Data from a new sample of 254 injured workers (M_{age} = 37.2 years, SD = 11.9, 53.5% men) was used to evaluate construct validity. The patients were seen for a half-day interdisciplinary assessment at the same rehabilitation center Switzerland during 2012. The YFQ was completed at home and amended if requested at intake. Four additional questionnaires used to measure construct validity were filled out at the rehabilitation center prior to the interdisciplinary assessment and therapeutic trial. First, the ten-item Neck Disability Index (NDI; [42]), which assesses ten aspects of disability such as neck pain intensity, personal care, lifting, reading, headaches, concentration, work, driving, sleeping, and recreation. Responses are given on a six-point scale ranging from *no disability* (0) to *total disability* (5). Higher scores indicate more severe self-reported disability. Then, self-reported anxiety and depression were measured using the Hospital Anxiety and Depression Scale (HADS; [43]). The HADS consists of seven-item subscales for anxiety and depression and responses are given on a four-point scale ranging from *best* (0) to *worst* (3). Total scores are calculated for each subscale (max. = 21) with higher scores indicating a more severe anxiety or depression. Last, self-efficacy was assessed using the Spinal Function Sort (SFS; [44]). The SFS assesses perceived ability to perform work tasks and activities of daily life and contains 50 drawings (e.g., put a glass bottle on the floor) with simple descriptions. The respondent is to rate his or her ability to carry out the task depicted in each drawing using a five-point Likert scale ranging from *able* (4) to *unable* (0). The SFS yields a single score (range 0–200). Based on the available literature, we expected that a substantial amount of the variance of the total YFQ-score would be positively predicted by self-reported pain, disability, as well as anxiety, depression and self-efficacy, yet for self-efficacy, we expected an inversed relationship. Moreover, due to its very specific work- and ADL-related tasks, it was expected that the self-efficacy measure would explain less variance in the YFQ than the disability, anxiety and depression measures.

Gender and age are widely acknowledged to be associated with self-reported pain experience (e.g., [1, 11]) and so these variables were included as control variables in the hierarchical linear regressions of *neck disability* and *mental health* on *YFQ score*.

Results

Phase 1: Development of a Preliminary Version of the YFQ

Included Clinical Experts

Two physicians licensed in physical medicine and rehabilitation, four clinical specialists with a background in occupational or physical therapy, a psychologist and one psychiatrist volunteered to participate.

Literature Search and Other Searches

The literature search led to items from the following well-established questionnaires which claim to measure fear avoidance beliefs (FABQ-D; [20]), pain attitudes (SOPA; [45]), pain anxiety (PASS; [21]), fear of movement (TSK-GV; [46]), impairment-beliefs (PAIRS; [47]), and pain catastrophizing beliefs (PCS; [16]). Additionally, items from the Yellow Flags list developed by Kendall et al. [9, 10] and other items used in clinical practice were gathered. Items about functional (dis-)ability during daily activities such as lifting, carrying, bending, or walking were added. If items from different questionnaires used similar items, consensus

was reached on one single item, or a blended version of the two items was developed [26].

First Expert Round

A total of 148 items were deemed relevant and served as a preliminary pool of items. Twenty-two items were derived from the established questionnaires, 22 new items emerged from blending items of the same established questionnaires, 94 items were deduced from clinical measures, which were collected based on patient statements during encounters and expert opinions.

Second Expert Round

Out of the 148 items included in the preliminary item pool, agreement was reached for 116 items grouped into 15 domains. These items included the domains Activity (10 items), Control (5), Distraction (6), Emotions (15), Harm (5), Diagnostics and Treatment (9), Pain attitudes (4), Goals (3), Blame (3), Self-prognosis (3), Social beliefs (5), Social support (16), Own work goals (14), and Beliefs about own work (18). Furthermore, consensus was achieved on a 5 point-Likert answering scale applied to all items ranging from *totally agree* (0) to *totally disagree* (4).

Phase 2: Psychometric Evaluations

From the pool of 116 YFQ-items, the Stanton and colleagues' criteria [28] were used to systematically select 48 items based on a quality score per item. Seven of the 48 items had more than 5% missing values. When checking the non-ignorable category of all missing values (omission of response was non-random), five further items were eliminated. Subsequently, the eight experts were asked for subjective qualitative judgments about the clinical relevance of the 48 items. They confirmed problems regarding five items with a high number of missing values, and classified six other items as lacking clinical relevance. The YFQ was thus reduced to 37 items (Appendix 1).

Confirmatory Factor Analyses Testing the Seven Versus the Eight-Domain Model of the YFQ

The Eight-Domain Model fit for the 37-item YFQ in subsample one demonstrated an acceptable fit based on values of χ^2/degrees of freedom, CFI and NFI and a good fit based on RMSEA. The data from sub-sample two revealed a slightly better fit, with the NFI indicating acceptable fit and χ^2/degrees of freedom, RMSEA and CFI indicating good fit (Table 1). Thus, the proposed Eight-Domain structure was supported by the data.

With regard to the Seven-Domain Model analysis for the 32-item YFQ performed in both subsamples, a good fit based on χ^2/degrees of freedom, RMSEA and CFI, and demonstrated an acceptable fit based on NFI. The Seven-Domain Model thus marked an overall better fit than the Eight-Domain Model. The result of the χ^2 difference test was significant in both sub-samples ($ps < .001$), confirming the superiority of the Seven-Domain Model compared to the Eight-Domain Model and further indicating that the domain pain beliefs might not necessarily constitute an independent domain within the set of yellow flags. Thus, the Seven-Domain YFQ consists of 32 items that represent the domains activity, co-morbidity, diagnosis beliefs, emotions, harm & blame, pain control, and work factors (Fig. 2). Means and standard deviations for the seven subscales are shown in Table 2.

Table 1 Fit Indices for different factorial models in structural equation modelling

	χ^2	df	χ^2/df	RMSEA	CFI	NFI
Random sub-sample 1 ($n = 333$)						
1. Null model	2646.65***	171	15.48	0.21	n.a.	n.a.
2. Eight-factor oblique model	270.37***	124	2.18	0.06	0.94	0.90
3. Seven-factor oblique model (domain pain beliefs excluded)	168.38***	98	1.72	0.05	0.97	0.92
Random sub-sample 2 ($n = 333$)						
1. Null model	3072.05***	171	18.00	0.23	n.a.	n.a.
2. Eight-factor oblique model	242.33***	124	1.95	0.05	0.96	0.92
3. Seven-factor oblique model (domain pain beliefs excluded)	156.68***	98	1.60	0.04	0.98	0.94

n.a. not applicable, χ^2 model chi-squared, *df* degrees of freedom, *RMSEA* root mean square error of approximation, *CFI* comparative fit index, *NFI* normed fit index, *SRMR* standardised root mean square residual

*$p < .05$, **$p < .01$, ***$p < .001$; all two-tailed

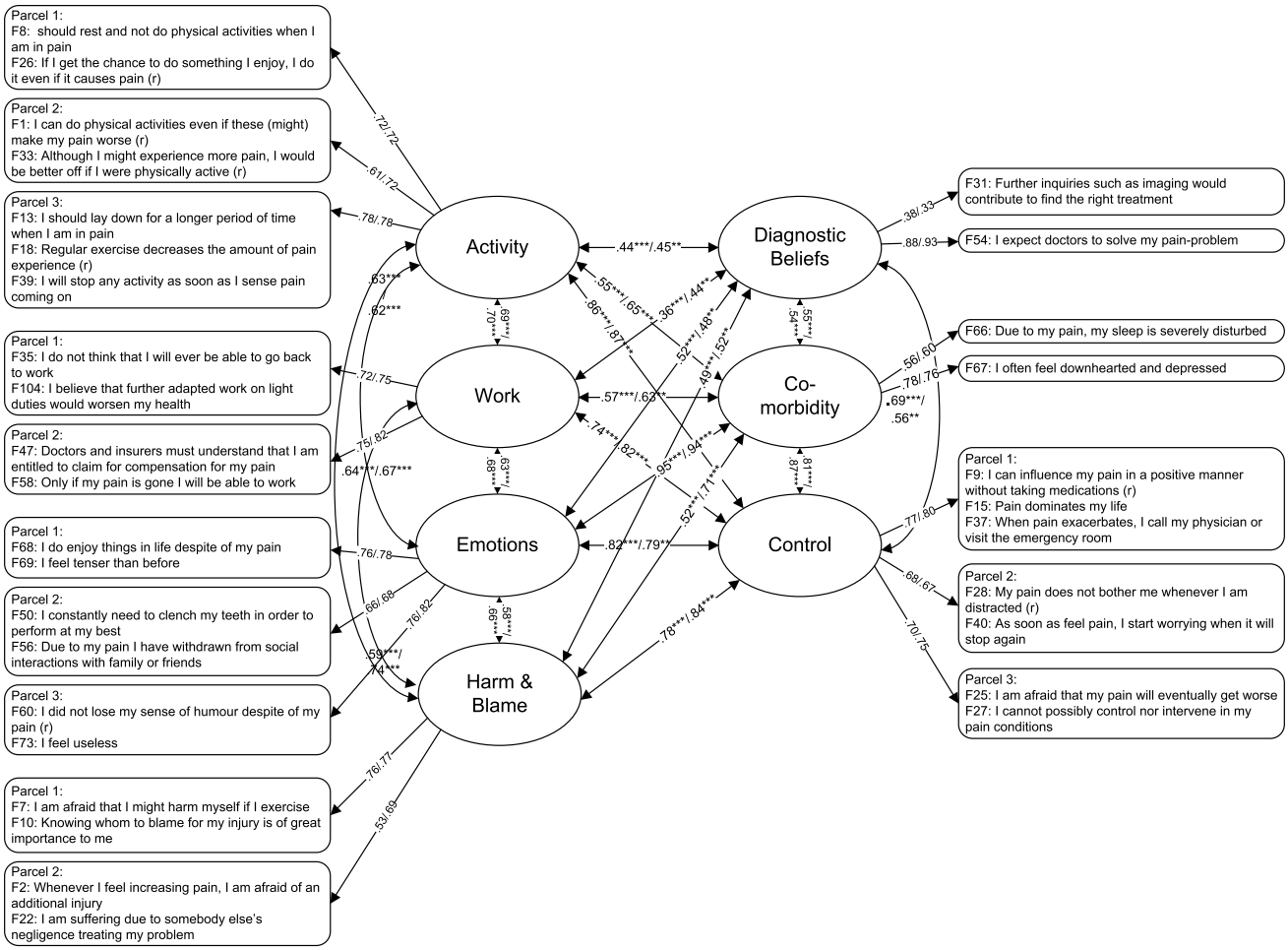

Fig. 2 Standardized path coefficients of the seven factorial confirmatory factor analysis using structural equation modelling and cross-validation in two random sub-samples. *Left path* coefficients for sub-sample 1; *right path* coefficients for sub-sample 2. All correlation coefficients differ significantly from zero ($p < .001$, two-tailed)

Phase 3: Validation Analyses

Reliability Analysis

Total YFQ score pointed out a high internal consistency of Cronbach's $\alpha = 0.91$. Internal consistency values for the domains were in majority of the cases satisfactory ranging from Cronbach's $\alpha = 0.49$–0.81, with exception of the domain diagnosis beliefs, where the values were Cronbach's $\alpha = 0.49$ and 0.57, respectively in both sub-samples (Table 3).

Construct Validity Analysis

The total YFQ score explained substantial proportions of the variance in disability and mental health indicators in hierarchical regression analyses ($\beta^2_{\text{neck disability}} = 42.25\%$, $\beta^2_{\text{anxiety}} = 47.61\%$, $\beta^2_{\text{depression}} = 44.89\%$ and $\beta^2_{\text{self-efficacy}} = 19.36\%$, see Table 4). No further variance was explained when adding the domain pain beliefs to the regression model. These results underline the YFQ's expected positive relationships with rehabilitation outcomes *musculoskeletal pain, anxiety* and *depression*, its expected negative relationship with *self-efficacy* and thus demonstrate its construct validity. The inclusion of the domain pain beliefs does not increase the construct validity of the YFQ.

Discussion

The results of the current study describe the step-wise approach to the development the YFQ, a multidimensional screening questionnaire. By using established questions, expert rounds, item-quality criteria scores [28] and statistical methods, a substantial reduction of the initial 148–37 items was achieved keeping its intended goal i.e., guiding clinical work with a questionnaire that reflects a variety of distinct yellow domains. The 37 items were assigned to 8

Table 2 Mean scores and standard deviations for the YFQ domains and rehabilitation outcome variables

	M	*SD*	*N*
Confirmatory factor analysis patient sample			
1. YFQ-activity	2.39	0.75	666
2. YFQ-work	1.31	0.91	549
3. YFQ-emotions	1.68	0.81	665
4. YFQ-harm & blame	1.54	0.92	600
5. YFQ-diagnosis beliefs	2.19	1.12	590
6. YFQ-co-morbidity	2.46	1.01	661
7. YFQ-pain control	2.22	0.85	662
8. YFQ-total	2.00	0.68	611
Construct validity analysis patient sample			
1. YFQ-activity	2.71	0.74	247
2. YFQ-work	0.96	0.93	195
3. YFQ-emotions	1.63	0.84	252
4. YFQ-harm & blame	1.23	0.93	185
5. YFQ-diagnosis beliefs	2.03	1.26	238
6. YFQ-co-morbidity	2.50	1.15	248
7. YFQ-pain control	2.08	0.95	250
8 YFQ-total	1.91	0.78	218
9. Neck disability	22.38	8.24	253
10. Anxiety	8.72	4.70	254
11. Depression	7.12	4.74	253
12. Self-efficacy	135.05	41.57	246

YFQ yellow flag questionnaire

domains, but indices of fit indicated that a model with 32 items divided into seven interrelated domains was a better fit than an Eight-Domain Model. The Seven-Domain YFQ further demonstrated good psychometric properties and construct validity. We conclude, therefore, that the shorter, Seven-Domain Model of the YFQ has high *internal item qualities, external item qualities* and *judgmental item qualities* [28]. Assessing the domain pain beliefs did not provide a better prediction of disability and distress indicators. This fact suggests that the crucial aspects of beliefs about the catastrophic consequences of pain are well summarized in the selected Seven-Domain Structure. Hence, it is unnecessary to increase the length of the questionnaire and define an eighth domain. Therefore, our main recommendation is for a use of the 32-item, Seven-Domain YFQ with items covering the domains activity, co-morbidity, diagnosis beliefs, emotions, harm & blame, pain control, and work factors.

Other screening tools, the STarT Back Tool (SBT; [12, 13]) with 9 items and the Orebro Musculoskeletal Pain Screening Questionnaire (OMPSQ; [14]) with 25 items are available. The SBT provides an overview and a first rough classification regarding a low, medium or high-risk group for ongoing disability. It was designed for clinicians in primary care. However, risk factors may change in the transition from

Table 3 Pearson correlations and internal consistencies for the YFQ domains and rehabilitation outcome variables

	1	2	3	4	5	6	7	8	9	10	11	12
1. YFQ-activity	(0.74/0.75)											
2. YFQ-work	0.50	(0.68/0.73)										
3. YFQ-emotions	0.43	0.51	(0.75/0.66)									
4. YFQ-harm & blame	0.40	0.56	0.51	(0.61/0.57)								
5. YFQ-diagnosis beliefs	0.38	0.43	0.38	0.47	(0.49/0.57)							
6. YFQ-Co-morbidity	0.54	0.52	0.65	0.52	0.37	(0.62/0.63)						
7. YFQ-pain control	0.64	0.62	0.72	0.58	0.49	0.68	(0.77/0.81)					
8 YFQ-total	0.72	0.77	0.78	0.75	0.69	0.81	0.87	(0.91/0.91)				
9. Neck disability	0.45	0.44	0.52	0.37	0.23	0.56	0.50	0.58	n.a.			
10. Anxiety	0.38	0.43	0.64	0.49	0.38	0.59	0.58	0.66	0.61	n.a.		
11. Depression	0.51	0.55	0.66	0.49	0.36	0.62	0.65	0.73	0.68	0.76	n.a.	
12. Self-efficacy	−0.47	−0.45	−0.41	−0.37	−0.22	−0.40	−0.46	−0.51	−0.63	−0.47	−0.53	n.a.

n.a. not applicable, *YFQ* Yellow Flag Questionnaire. Numbers in parentheses are values of Cronbach's alpha for sub-samples 1 and 2 (left- and right-hand sides, used in the confirmatory factor analysis and convergent validity test respectively). All correlation coefficients differ significantly from zero ($p < .001$, two-tailed)

Table 4 Summary of hierarchical regression of YFQ on (A) neck disability and anxiety indicators and (B) depression, and self-efficacy indicators

Predictor	Neck disability ($N=217$)				Anxiety ($N=217$)			
	b	*SE b*	β	R_{total}^2	*b*	*SE b*	β	R_{total}^2
(A)								
Step 1				.00				.00
Age	0.05	0.05	0.07		0.03	0.03	0.09	
Sex (1 = female, 2 = male)	0.20	1.15	0.01		−0.80	0.64	−0.09	
Step 2				.34***				.48***
Age	0.06	0.04	0.09		0.05	0.02	0.12*	
Sex (1 = female, 2 = male)	−1.38	0.94	−0.08		−1.84	0.47	−0.19***	
Seven-domain YFQ	6.31	0.59	0.60***		4.13	0.30	0.70***	
Step 3				.34***				.47***
Age	0.06	0.04	0.09		0.05	0.02	0.12*	
Sex (1 = female, 2 = male)	−1.38	0.94	−0.08		−1.84	0.47	−0.19***	
Seven-domain YFQ	6.89	0.94	0.65***		4.09	0.47	0.69***	
YFQ domain pain beliefs	−0.62	0.79	−0.07		0.04	0.40	0.01	
Predictor	Depression ($N=216$)				Self-efficacy ($N=213$)			
	b	*SE b*	β	R_{total}^2	*b*	*SE b*	β	R_{total}^2
(B)								
Step 1				.01				.00
Age	0.03	0.03	0.08		−0.25	0.25	−0.07	
Sex (1 = female, 2 = male)	0.75	0.64	0.08		6.03	5.94	0.07	
Step 2				.53***				.27***
Age	0.04	0.02	0.11*		−0.33	0.21	−0.09	
Sex (1 = female, 2 = male)	−0.33	0.44	−0.04		12.95	5.11	0.15*	
Seven-domain YFQ	4.34	0.28	0.73***		−28.85	3.21	−0.53***	
Step 3				.53***				.28***
Age	0.04	0.02	0.11*		−0.33	0.21	−0.09	
Sex (1 = female, 2 = male)	−0.33	0.44	−0.04		12.99	5.11	0.15*	
Seven-domain YFQ	3.98	0.44	0.67***		−24.10	5.18	−0.44***	
YFQ domain pain beliefs	0.39	0.38	0.08		−5.20	4.44	−0.11	

b non-standardised regression coefficient, *SE* standard error of non-standardised regression coefficient, *β* standardised regression coefficient, *YFQ* yellow flag questionnaire

*$p<.05$, **$p<.01$, ***$p<.001$; all two-tailed

acute to chronic phase. The SBT lacks validity in addressing the risk factors in chronic cases, likewise is a SBT-based treatment for that population. The second screening questionnaire, the OMPSQ, was developed for early identification of yellow flags in patients risking the development of pain-related work disability [48]. The validity of the OMPSQ in a population with chronic pain has still to be established since a recent study did not support the factor structure of the OMPSQ in worker with chronic MSK [15].

Another difference between the OMPSQ and the YFQ lies in the structures. A minimum of two questions per domain cover the YFQ's seven basic domains related to potential barriers to RTW, whereas OMPSQ refers to an overall cut-off score of 105 for predicting a positive or negative outcome. Unlike OMPSQ, the YFQ provides indications which of the 7 domains should be addressed treating injured workers. Some differences occur also in the thematic priorities since the OMPSQ includes five of six basic factors related to pain in the short-form [48]: self-perceived function, pain experience, distress, fear-avoidance beliefs, and return to work expectancy. While there is some overlap with the factors of the OMPSQ, disability-related cognitions, emotions and behaviors are at the core of YFQ rather than the pain-experience. According to a recent systematic review, multidisciplinary treatments display rather poor efficacy regarding pain-improvement over time, and acceptable efficacy in improvement of disability [49]. Whether the YFQ will be useful in accurately measure changes in disability

after multidisciplinary treatments should be established in future studies.

The "Decade of the Flags" Working Group commented on the challenges in developing screening tools [3]. Their argument that a screening instrument is never 100% accurate and the conclusion to create screenings with high sensitivity, but accept low specificity in order to minimize the chances of missing a positive case supports the effort to develop the multidimensional Flag-based construct of the YFQ with a broader spectrum of potential psychological risk factors. Furthermore, the challenge of developing a universal screening instrument for multiple purposes such as measuring low-term pain outcomes, disability and work absenteeism has been highlighted in a recent systematic review [50]. While the OMPSQ performed well in discriminating work absenteeism, the same instrument performed acceptable and poorly in disability and pain outcomes, respectively [50].

Practical Implications

The intention of developing the YFQ emerged from the practical need for a self-reported measure that would allow capturing a range of flag-based risk factors with a single questionnaire. While the YFQ has anecdotally shown clinical utility over the last decade, this study provides first evidence about the measurement properties of the instrument. Within its seven domains, the YFQ includes many relevant risk factors that would otherwise need to be appraised with multiple questionnaires. The current electronic version of the YFQ visualizes the total scores as well as sub-scores of each domain in a bar diagram (from 0 to 100% total score). Therefore, it's expected that the YFQ would guide conversations between health care professionals and patients. Ultimately, domains with elevated scores may be addressed by individually tailored interventions such as education and cognitive-behavioral therapy combined with physical therapy [11]. In addition, the YFQ could be implemented to measure effects of education programs aimed to alter maladaptive attitudes and beliefs. More qualitative and quantitative research is needed to understand how the YFQ assist the rehabilitation process from a patient and provider perspective.

Strengths and Limitations

One strength of the study is the large pool of well-established clinical yellow flags that was substantially reduced in a step-wise procedure to develop a questionnaire consisting of the minimum number of items required to assess recovery obstacles in patients with chronic MSK. This bottom-up approach was complemented by the top-down development of new items using expert rounds inspired by the Delphi method.

This study has limitations. First, although the reliability of the total YFQ score is high, the reliabilities of the domains diagnosis beliefs, harm and blame, and co-morbidity are unsatisfactory. This is due to two reasons: First, Cronbach's Alpha values are sensitive to the number of included items [40]. Both domains, diagnosis beliefs and co-morbidity, consist of only two items each. Thus, considering the minimal number of two included items, both Alpha values score rather high in both domains [41]. Similar explanations need to be taken into consideration regarding the domain harm & blame, which is thematically divided into two topics: harm and blame. Again, Alpha values refer to the internal consistency, which is limited in two topics, however closely related. Second, construct validity was used with a limited set of questionnaires. There is a need to validate the seven domains with additional scales in further studies. Third, the results of this study apply only to patients with MSK who were referred to a specialized clinic. More studies are needed to validate the YFQ in other contexts, such as in primary care and private practice. Further work should also include structural calculations to define standardized test norms for comparison purposes. While an arbitrary cut off score of 50% (YFQ score of the individual/YFQ max score) for a "high" or "low" risk patient is suggested by the developers, scientific validation of that cut-off value with external criteria is lacking, as is normative data to compare individual scores that might have clinical value in terms of guiding treatment and allocating resources.

Conclusions

The YFQ, a multidimensional screening scale is recommended for use in assessing psychosocial beliefs of workers with chronic MSK pain. Further evaluation of the measurement properties such as the test–retest reliability, responsiveness and prognostic validity is warranted.

Acknowledgements The authors thank the patients and clinicians of the Department of Work Rehabilitation in the Rehabilitation Clinic in Bellikon for the help in the development of the Yellow Flag Questionnaire. Furthermore, the authors thank Stefan Kälin for his help in item analysis.

Funding Part of the study was funded by the Swiss Accident Insurance Fund, SUVA (Schweizerische Unfallversicherungsanstalt).

Author Contributions Maurizio Trippolini and Michael Oliveri were project leaders, conceived the design of the study, provided funding, and data collection. Livio Terribilini performed data analysis and wrote the first draft of the manuscript. Cornelia Rolli Salathé and Achim Elfering structured the ideas, offered statistical support, performed psychometric analyses, and wrote the manuscript. All authors were involved in data interpretation, revised the manuscript, and gave final approval of the manuscript.

Compliance with Ethical Standards

Conflict of interest All authors declare that they have no financial or non-financial competing of interests related to this study.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Patients were only included in the study if they had consented to the use of their data for research purposes. The ethical committee of Aargau canton, Switzerland approved the study.

Appendix 1

The Yellow Flag Questionnaire (YFQ)

Instructions: *Please read each statement and indicate how much you agree. Please choose only one answer per question*.

☐ Strongly Agree.
☐ Somewhat Agree.
☐ Unsure.
☐ Somewhat Disagree.
☐ Strongly Disagree.

Activity

F1. I can do physical activities even if these (might) make my pain worse *[r]*.

F8. I should rest and not do physical activities when I am in pain.

F13. I should lay down for a longer period of time when I am in pain.

F18. Regular exercise decreases the amount of pain experience *[r]*.

F26. If I get the chance to do something I enjoy, I do it even if it causes pain *[r]*.

F33. Although I might experience more pain, I would be better off if I were physically active *[r]*.

F39. I will stop any activity as soon as I sense pain coming on.

Work

F35. I do not think that I will ever be able to go back to work.

F47. Doctors and insurers must understand that I am entitled to claim for compensation for my pain.

F58. Only if my pain is gone I will be able to work.

F104. I believe that further adapted work on light duties would worsen my health.

Emotions

F50. I constantly need to clench my teeth in order to perform at my best.

F56. Due to my pain I have withdrawn from social interactions with family or friends.

F60. I did not lose my sense of humour despite of my pain *[r]*.

F68. I do enjoy things in life despite of my pain.

F69. I feel tenser than before.

F73. I feel useless.

Harm and Blame

F2. Whenever I feel increasing pain, I am afraid of an additional injury.

F7. I am afraid that I might harm myself if I exercise.

F10. Knowing whom to blame for my injury is of great importance to me.

F22. I am suffering due to somebody else's negligence treating my problem.

Diagnostic Beliefs

F31. Further inquiries such as imaging would contribute to find the right treatment.

F54. I expect doctors to solve my pain-problem.

Co-morbidity

F66. Due to my pain, my sleep is severely disturbed.

F67. I often feel downhearted and depressed.

Control

F9. I can influence my pain in a positive manner without taking medications *[r]*.

F15. Pain dominates my life.

F25. I am afraid that my pain will eventually get worse.

F27. I cannot possibly control nor intervene in my pain conditions.

F28. My pain does not bother me whenever I am distracted *[r]*.

F37. When pain exacerbates, I call my physician or visit the emergency room.

F40. As soon as feel pain, I start worrying when it will stop again.

Notes: F-codes refer to the numbering in the original sample of 116 questionnaire items; [r] indicates a reverse-scored item.

References

1. Elfering A, Mannion AF. Epidemiology and risk factors of spinal disorders. In: Boos N, Aebi M, editors. Spinal disorders – fundamentals of diagnosis and treatment. Berlin: Springer; 2008. pp. 153–306.
2. Main CJ, Williams ACdC. ABC of psychological medicine: musculoskeletal pain. BMJ. 2002;325(7363):534–537.
3. Nicholas MK, Linton SJ, Watson PJ, Main CJ. Early identification and management of psychological risk factors ("yellow flags") in patients with low back pain: a reappraisal. Phys Ther. 2011;91(5):737–753.
4. Melloh M, Elfering A, Egli Presland C, Roeder C, Barz T, Rolli Salathé C, et al. Identification of prognostic factors for chronicity in patients with low back pain: a review of screening instruments. Int Orthop. 2009;33(2):301–313.
5. Krause N, Frank JW, Dasinger LK, Sullivan TJ, Sinclair SJ. Determinants of duration of disability and return-to-work after work-related injury and illness: challenges for future research. Am J Ind Med. 2001;40(4):464–484.
6. Iles RA, Davidson M, Taylor NF. Psychosocial predictors of failure to return to work in non-chronic non-specific low back pain: a systematic review. Occup Environ Med. 2008;65(8):507–517.
7. Iakova M, Ballabeni P, Erhart P, Seichert N, Luthi F, Dériaz O. Self perceptions as predictors for return to work 2 years after rehabilitation in orthopedic trauma inpatients. J Occup Rehabil. 2012;22(4):532–540.
8. Young AE, Wasiak R, Gross DP. Recurrence of work-related low back pain and disability: association between self-report and workers' compensation data. Spine. 2013;38(26):2279–2286.
9. Kendall NA. Psychosocial approaches to the prevention of chronic pain: the low back paradigm. Best Pract Res Clin Rheumatol. 1999;13(3):545–554.
10. Kendall NAS, Linton SJ, Main C. Guide to assessing psychosocial yellow flags in acute low back pain: risk factors for long-term disability and work loss. Wellington, New Zealand: ACCatNZG Group; 2004.
11. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. Lancet. 2017;389(10070):736–747.
12. Hill JC, Dunn KM, Lewis M, Mullis R, Main CJ, Foster NE, et al. A primary care back pain screening tool: identifying patient subgroups for initial treatment. Arthrit Care Res. 2008;59(5):632–641.
13. Hill JC, Vohora K, Dunn KM, Main CJ, Hay EM. Comparing the STarT back screening tool's subgroup allocation of individual patients with that of independent clinical experts. Clin J Pain. 2010;26(9):783–787.
14. Linton SJ, Boersma K. Early identification of patients at risk of developing a persistent back problem: the predictive validity of the Örebro Musculoskeletal Pain Questionnaire. Clin J Pain. 2003;19(2):80–86.
15. Soer R, Vroomen P, Stewart R, Coppes M, Stegeman P, Dijkstra P, et al. Factor analyses for the Örebro Musculoskeletal Pain Questionnaire for working and nonworking patients with chronic low back pain. Spine J. 2017;17(4):603–609.
16. Sullivan MJL, Bishop SR, Pivik J. The Pain Catastrophizing Scale: development and validation. Psychol Assessment. 1995;7(4):524–532.
17. Lorig K, Chastain RL, Ung E, Shoor S, Holman HR. Development and evaluation of a scale to measure perceived self-efficacy in people with arthritis. Arthritis Rheum. 1989;32(1):37–44.
18. Robinson ME, Riley JL III, Myers SD, Sadler IJ, Kvaal SA, Geisser ME, et al. The Coping Strategies Questionnaire: a large sample, item level factor analysis. Clin J Pain. 1997;13(1):43–49.
19. Flor H, Behle DJ, Birbaumer N. Assessment of pain-related cognitions in chronic pain patients. Behav Res Ther. 1993;31(1):63–73.
20. Staerkle R, Mannion AF, Elfering A, Junge A, Semmer NK, Jacobshagen N, et al. Longitudinal validation of the Fear-Avoidance Beliefs Questionnaire (FABQ) in a Swiss-German sample of low back pain patients. Eur Spine J. 2004;13(4):332–340.
21. McCracken LM, Zayfert C, Gross RT. The Pain Anxiety Symptoms Scale: development and validation of a scale to measure fear of pain. Pain. 1992;50(1):67–73.
22. Elfering A, Mueller U, Rolli Salathé C, Tamcan O, Mannion AF. Pessimistic back beliefs and lack of exercise: a longitudinal risk study on shoulder, neck, and back pain. Psychol Health Med. 2015;20(7):767–780.
23. Sullivan MJ, Yakobov E, Scott W, Tait R. Perceived injustice and adverse recovery outcomes. Psychol Inj Law. 2014;7(4):325–334.
24. Baxendale S. When less is more. Data reduction in the prediction of postoperative outcome. Epilepsy Behav. 2014;31:219. doi:10.1016/j.yebeh.2013.11.004.
25. Galesic M, Bosnjak M. Effects of questionnaire length on participation and indicators of response quality in a web survey. Public Opin Quart. 2009;73(2):349–360.
26. Dalkey N, Helmer O. An experimental application of the Delphi method to the use of experts. Manag Sci. 1963;9(3):458–467.
27. Barbour RS. The case for combining qualitative and quantitative approaches in health service research. J Health Serv Res Policy. 1999;4(1):39–43.
28. Stanton JM, Sinar EF, Balzer WK, Smith PC. Issues and strategies for reducing the length of self-report scales. Pers Psychol. 2002;55(1):167–194.
29. McHorney CA, Bricker DE, Robbins JA, Kramer AE, Rosenbek JC, Chignell KA. The SWAL-QOL outcomes tool for oropharyngeal dysphagia in adults: II. Item reduction and preliminary scaling. Dysphagia. 2000;15(3):122–133.
30. Boyle GJ. Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? Pers Indiv Differ. 1991;12(3):291–294.
31. Smith PC, Stanton JM. Perspectives on the measurement of job attitudes: the long view. Hum Resour Manage R. 1999;8(4):367–386.
32. Gabel CP, Melloh M, Yelland M, Burkett B, Roiko A. Predictive ability of a modified Örebro Musculoskeletal Pain Questionnaire in an acute/subacute low back pain working population. Eur Spine J. 2011;20(3):449–457.
33. Paik MC. Non-ignorable missingness in matched case–control data analyses. Biometrics. 2004;60(2):306–314.
34. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–592.
35. Heitjan DF, Basu S. Distinguishing "missing at random" and "missing completely at random". Am Stat. 1996;50(3):207–213.
36. Bandalos DL. The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. Struct Equ Modeling. 2002;9(1):78–102.
37. Little TD, Cunningham WA, Shahar G. To parcel or not to parcel: exploring the question, weighing the merits. Struct Equ Modeling. 2002;9(2):151–173.
38. Bagozzi RP, Edwards JR. A general approach for representing constructs in organizational research. Organ Res Methods. 1998;1(1):45–87.

39. Hooper D, Coughlan J, Mullan MR. Structural equation modeling: Guidelines for determining model fit. Electron J Business Res Methods. 2008;6(1):53–60.
40. Moosbrugger H, Schermelleh-Engel K. Exploratorische (EFA) und Konfirmatorische Faktorenanalyse (CFA). In: Moosbrugger H, Kelava A, editors. Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer; 2007. pp. 307–24.
41. Körner A, Geyer M, Brähler E. Das NEO-Fünf-Faktoren Inventar (NEO-FFI): Validierung anhand einer deutschen Bevölkerungsstichprobe. Diagnostica. 2002;48(1):19–27.
42. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manip Physiol Ther. 1991;14(7):409–415.
43. Herrmann-Lingen C, Buss U, Snaith RP. Hospital Anxiety and Depression Scale - Deutsche Version (HADS-D). Manual. 3rd ed. Bern: Hans Huber; 2011.
44. Matheson LN, Matheson ML, Grant J. Development of a measure of perceived functional ability. J Occup Rehabil. 1993;3(1):15–30.
45. Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: a comparison of six methods. Pain. 1986;27(1):117–126.
46. Rusu AC, Kreddig N, Hallner D, Hülsebusch J, Hasenbring MI. Fear of movement/(Re) injury in low back pain: confirmatory validation of a German version of the Tampa Scale for Kinesiophobia. BMC Musculoskelet Disord. 2014;15(1):280–289.
47. Riley JF, Ahern DK, Follick MJ. Chronic pain and functional impairment: Assessing beliefs about their relationship. Arch Phys Med Rehab. 1988;69(8):579–582.
48. Linton SJ, Nicholas M, MacDonald S. Development of a short form of the Örebro Musculoskeletal Pain Screening Questionnaire. Spine. 2011;36(22):1891–1895.
49. Kamper SJ, Apeldoorn AT, Chiarotto A, Smeets RJ, Ostelo RW, Guzman J, et al. Multidisciplinary biopsychosocial rehabilitation for chronic low back pain. Cochrane Library. 2014;(9):CD000963. doi:10.1002/14651858.CD000963.pub3.
50. Karran EL, McAuley JH, Traeger AC, Hillier SL, Grabherr L, Russek LN, et al. Can screening instruments accurately determine poor outcome risk in adults with recent onset low back pain? A systematic review and meta-analysis. BMC Med. 2017;15(1):13.