The Journal of Rheumatology Volume 46, no. 10

Considerations for Evaluating and Recommending Worker Productivity Outcome Measures: An Update from the OMERACT Worker Productivity Group

Suzanne M.M. Verstappen, Diane Lacaille, Annelies Boonen, Reuben Escorpizo, Catherine Hofstetter, Ailsa Bosworth, Amye Leong, Sarah Leggett, Monique A.M. Gignac, Johan K. Wallman, Marieke M. Ter Wee, Florian Berghea, Maria Agaliotis, Peter Tugwell and Dorcas Beaton

J Rheumatol 2019;46;1401-1405
http://www.jrheum.org/content/46/10/1401

Considerations for Evaluating and Recommending Worker Productivity Outcome Measures: An Update from the OMERACT Worker Productivity Group

Suzanne M.M. Verstappen, Diane Lacaille, Annelies Boonen, Reuben Escorpizo, Catherine Hofstetter, Ailsa Bosworth, Amye Leong, Sarah Leggett, Monique A.M. Gignac, Johan K. Wallman, Marieke M. Ter Wee, Florian Berghea, Maria Agaliotis, Peter Tugwell, and Dorcas Beaton

ABSTRACT. ***Objective.*** The Outcome Measures in Rheumatology (OMERACT) Worker Productivity Group continues efforts to assess psychometric properties of measures of presenteeism.

Methods. Psychometric properties of single-item and dual answer multiitem scales were assessed, as well as methods to evaluate thresholds of meaning.

Results. Test-retest reliability and construct validity of single item global measures was moderate to good. The value of measuring both degree of difficulty and amount of time with difficulty in multi-items questionnaires was confirmed. Thresholds of meaning vary depending on methods and external anchors applied.

Conclusion. We have advanced our understanding of the performance of presenteeism measures and have developed approaches to describing thresholds of meaning. (First Release May 15 2019; J Rheumatol 2019;46:1401–5; doi:10.3899/jrheum.181201)

Key Indexing Terms:
OMERACT PSYCHOMETRIC PROPERTIES MINIMUM IMPORTANT DIFFERENCE
DUAL PRESENTEEISM SCALE PATIENT ACCEPTABLE STATE PRESENTEEISM

From the Arthritis Research UK Centre for Epidemiology, Centre for Musculoskeletal Research, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre; UK National Institute for Health Research (NIHR) Manchester Biomedical Research Centre, Manchester University Hospitals National Health Service (NHS) Foundation Trust, Manchester Academic Health Science Centre, Manchester; Arthritis Research UK/Medical Research Council (MRC) Centre for Musculoskeletal Health and Work, University of Southampton, Southampton; National Rheumatoid Arthritis Society (NRAS), Maidenhead, UK; Department of Medicine, and the Division of Rheumatology, University of British Columbia, Vancouver; Arthritis Research Canada, Richmond, British Columbia; Institute for Work & Health; University of Toronto, Toronto; Division of Rheumatology, Department of Medicine, and School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa; Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada; Division of Rheumatology, Maastricht University Medical Center, Care and Public Health Research Institute (CAPHRI), Maastricht; Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, the Netherlands; Department of Rehabilitation and Movement Science, University of Vermont, Burlington, Vermont, USA; Swiss Paraplegic Research, Nottwil, Switzerland; Lund University, Skåne University Hospital, Department of Clinical Sciences Lund, Rheumatology, Lund, Sweden; Carol Davila University of Medicine, Bucharest, Romania; School of Public Health and Community Medicine, University of New South Wales, Kensington, Australia.

S.M. Verstappen, PhD, MSc, Arthritis Research UK Centre for Epidemiology, Centre for Musculoskeletal Research, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, and NIHR Manchester Biomedical Research Centre, Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, and Arthritis Research UK/MRC Centre for Musculoskeletal Health and Work, University of Southampton; D. Lacaille, MDCM, FRCPC, MHSc, Department of Medicine, and the Division of Rheumatology, Department of Medicine, University of British Columbia, and Arthritis Research Canada; A. Boonen, MD, PhD, Division of Rheumatology, Maastricht University Medical Center, CAPHRI; R. Escorpizo, BSc, MSc, DPT, Department of Rehabilitation and Movement Science, University of Vermont, and Swiss Paraplegic Research; C. Hofstetter, OMERACT Patient Research Partner; A. Bosworth, MBE, OMERACT Patient Research Partner, NRAS; A. Leong, OMERACT Patient Research Partner; S. Leggett, MSc, Arthritis Research UK Centre for Epidemiology, Centre for Musculoskeletal Research, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre; M.A. Gignac, PhD, Institute for Work & Health, and University of Toronto; J.K. Wallman, MD, PhD, Lund University, Skåne University Hospital, Department of Clinical Sciences Lund, Rheumatology; M.M. Ter Wee, PhD, MSc, Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Epidemiology and Biostatistics, and Department of Rheumatology, Amsterdam Infection and Immunity Institute, Amsterdam Public Health; F. Berghea, MD, PhD, Carol Davila University of Medicine; M. Agaliotis, PhD, MSc, Australian Institute of Health Management Services, University of Tasmania; P. Tugwell, MD, MSc, Division of Rheumatology, Department of Medicine, and School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, and Clinical Epidemiology Program, Ottawa Hospital Research Institute; D. Beaton, PhD, Institute for Work & Health, and University of Toronto.

Address correspondence to Dr. S.M.M. Verstappen, Arthritis Research UK Centre for Epidemiology, Centre for Musculoskeletal Research, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK.
E-mail: Suzanne.Verstappen@manchester.ac.uk

Accepted for publication March 22, 2019.

Quantifying restrictions in worker participation, including absenteeism, sick leave, and presenteeism (i.e., reduced productivity because of ill health), is an important outcome

from a patient's perspective and is increasingly seen as a health outcome to target for improvement. People with rheumatic and musculoskeletal diseases (RMD) can experience variable levels of presenteeism and absenteeism depending on their health status, job demands, or other personal or environmental contextual factors[1].

During the last 8 years, the Outcome Measures in Rheumatology (OMERACT) Worker Productivity Group has evaluated available measures to assess worker productivity loss, initiated new research to fill in knowledge gaps regarding psychometric properties, and appraised these measures against the OMERACT Filter 2.1[1,2,3,4]. Based on a review of available instruments in the literature, we had a mandate to move forward with 6 candidate measures (4 single-item global and 2 multiitem measures) to assess presenteeism[2]: Worker Productivity Scale-Arthritis (WPS-A)[5], Work Productivity and Activity Impairment Questionnaire (WPAI)[6], Work Ability Index (WAI)[7], Quality and Quantity (QQ) questionnaire[8], Workplace Activity Limitations Scale (WALS)[9], and the modified Work Limitation Questionnaire-25 ($WLQ\text{-}25_{PDmod}$)[10]. These can be organized into a taxonomy of 4 different types of worker productivity measures, which sit against the background of contextual factors (Figure 1). At OMERACT 12, we received support (> 70% consensus) that $WLQ\text{-}25_{PDmod}$, WALS, and WAI had enough OMERACT Filter evidence available and we are conducting ongoing research for these measures for future endorsement, while also continuing to monitor QQ. Since OMERACT 12, we have progressed in our research across the following 4 workstreams: (1) collating further evidence about reliability, content/construct validity of global (i.e., single-item) measures of presenteeism and supplementing information on WPS-rheumatoid arthritis (RA) and WPAI, which were previously endorsed; (2) evaluation of psychometric of dual answer scales of 2 validated multiitem measures; and (3) determination of patient acceptable state (PAS) and the minimal important difference (MID) of presenteeism measures; and (4) contextual factors.

MATERIALS AND METHODS

Special Interest Group (SIG) OMERACT 2018. At OMERACT 14, we presented an update of our work on the first 3 workstreams. Attendees at our SIG included patients (n = 4), clinicians (n = 7), 1 fellow, and others (e.g., methodologists, industry, n = 5). Important questions were discussed with participants during breakout sessions, including:

1) Global measures: Based on the results presented (reliability, cross-cultural differences, construct validity), what would be your preferred global measure and why?

2) Multiitem measure: Based on the context of your research, or your experience as a patient, what do you think are the advantages and drawbacks of using answers that assess both the degree of difficulty and the amount of time with difficulty?

3) PAS/MID: (1) How best to manage MID thresholds and (2) do you agree with the need to report multiple MID/PAS thresholds?

Ethics approval was obtained for individual studies and all patients provided written informed consent [Making It Work trial: University of British Columbia Research Ethics Board (H11-03527); the EULAR-PRO (European League Against Rheumatism — Patient Reported Outcomes) study obtained overall ethical approval from National Research Ethics

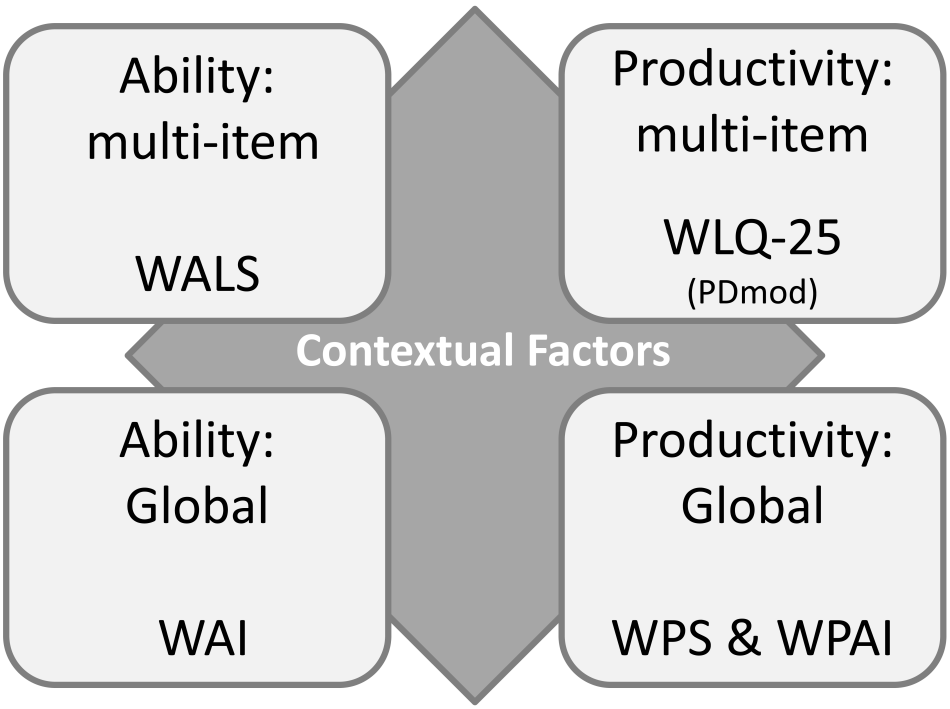

Figure 1. Organization into a taxonomy of 4 different types of work productivity measures. WALS: Workplace Activity Limitations Scale; WLQ-25 PDmod: Work Limitations Questionnaire with modified physical demands scale; WAI: Work Ability Index; WPS: Arthritis-specific Worker Productivity Scale; WPAI: Work Productivity and Activity Impairment Questionnaire.

Service Committee NW–Greater Manchester (12/NW/0172), and from each participating center according to national guidelines].

Global measures. To address the meaning of at-work productivity loss measures from a patient's perspective in different cultures, we conducted the international EULAR-PRO study to assess presenteeism in patients with inflammatory arthritis (IA) or osteoarthritis. The results of phase I have been published and show fair to excellent test-retest reliability [ICC for Health Productivity Questionnaire (HPQ; question C; 0.59) to WPS-RA (0.78)][11]. In-depth cognitive debriefing interviews revealed variation in how participants interpreted some of the constructs among the 5 measures, especially regarding "performance" in the HPQ scale, which was a term used in sport and theatre but not related to work for participants from Romania and Sweden[12]. For most participants (~70%), a recall period of 7 days up until a month would be a good reflection of the effect their health has on work. Phase II is an international observational cohort study (n = 8 countries) to further test psychometric properties. Preliminary results of baseline data on construct validity were presented during the SIG and show moderate to good construct validity (Table 1)[13]. During the break-out session, SIG attendees agreed that a recall period of 1 day was not representative, although they thought a recall period of a month might be too long. Other discussion points included wording of anchors (e.g., normal). Further, participants highlighted the difficulty in answering and interpreting disease-specific scales, because of the complexity of many rheumatic diseases, and preferred a generic scale.

Multiitem measures. How to best measure presenteeism using multiitem scales remains challenging. The WLQ and WALS are frequently used, but participants' feedback expressed concern about the constructs measured by each instrument. The WLQ measures the amount of time people are limited, but not the extent to which they are limited. This was perceived as a drawback by patients who felt it misses an important part of their experience and by researchers interested in evaluating presenteeism as a health state. In contrast, the WALS measures the extent of limitation but not time. This was a drawback to health economists, because of difficulty assigning cost. To evaluate psychometric properties encompassing both concepts, items from each measure (WALS and WLQ) were offered both time and difficulty response keys (dual answer keys).

Baseline and 6-month data were used from a Canadian randomized controlled trial (Making It Work Program) of an employment intervention including patients with IA (n = 364)[14,15]. The psychometric properties of the measures were first evaluated with the 2 answer keys analyzed separately (i.e., without combining results)[16]. Answers from the dual answer keys were then combined into a single score, obtained by (1) multiplying or (2) adding scores of difficulty and time answer keys at the item level[17]. No significant differences were observed between additive and multiplicative models. High correlation (≥ 0.8) between difficulty and time was found in only 2/12 WALS items and 11/25 WLQ, justifying the need for dual answer keys. High internal consistency (i.e., ≥ 0.7) was found for WALS and all WLQ subscales for both answer keys analyzed separately and combined (except WLQ-Physical Demands)[16]. As *a priori* hypothesized, moderate correlations were observed between original answer keys, or combined scores, of WLQ subscales and WALS with measures assessing similar concepts [WPAI; work instability scale (WIS; congruent validity)]. During the SIG, all agreed that dual answer keys provided additional value. Patient representatives uniformly said they felt that asking both degree and time with difficulty better reflected their experience, and that asking time alone would miss an important concept. The main concern raised was the length and complexity of the questionnaire with both answer keys. Other issues raised included concern about the 2-week recall period and descriptors for time options (considered difficult to answer by patients), and concern about percentage of time attributed to descriptor (e.g., some of the time = 50% of the time).

Thresholds of meaning for worker productivity measures. Thresholds of meaning are benchmarks for scores (e.g., PAS of pain) or change in scores [e.g., minimal threshold for change to be important (MCID)] that aid in the interpretation at an individual patient level. Recently, Copay, *et al* has demonstrated that there are considerable differences in MCID thresholds depending on the anchor or method[18]. At OMERACT in 2018, our focus was on dealing with these differences. As a group, we had reviewed the literature on these attributes and decided on best methods for their determination. In doing so, we emphasized the pivotal role of a meaningful anchor that becomes a gold standard for threshold determination, and the methods used to determine the actual cutoff. We fielded several anchors and provided several analytic approaches to each, allowing us to see the differences in values obtained, which also led to differences in the proportion deemed to be "improved" or "in an acceptable state" (Figure 2).

During our SIG, most of the attendees agreed that we will need to work with a range of MID values. There are also new developments and approaches in reporting results for thresholds, such as cumulative distribution function[19]. The various thresholds for MCID are highlighted with a vertical line on the same graph and demonstrate not only the proportion

Table 1. Construct validity of 4 global measures of presenteeism (WPAI, WPS-A, WAI, QQ) with the multiitem presenteeism measures WALS and patient-reported health outcome measures. Values are r.

Measures	WPAI	WPS-A	WAI	QQ-Quantity	QQ-Quality	QQ-Total*
Global presenteeism measures						
WPAI	1.0					
WPS-A	0.83	1.0				
WAI	–0.65	–0.62	1.0			
QQ-Quantity	–0.58	–0.53	0.60	1.0		
QQ-Quality	–0.52	–0.49	0.58	0.75	1.0	
QQ-Total	–0.60	–0.56	0.63	0.95	0.88	1.0
Multiitem presenteeism measure						
WALS	0.65	0.64	–0.55	–0.50	–0.49	–0.54
Health-related patient reported outcomes						
VAS general health	0.54	0.51	–0.43	–0.39	–0.36	–0.42
EQ-5D	–0.54	–0.54	0.48	0.37	0.39	0.42
HAQ	0.57	0.58	–0.52	–0.40	–0.41	–0.45

*For QQ-Total, the quality and quantity score are multiplied, resulting in a score between 0 and 100. WPS-A: Worker Productivity Scale-Arthritis (0 = no interference to 10 = complete interference); WPAI: Work Productivity and Activity Impairment Questionnaire (score 0 = condition no effect on work to 10 = condition completely prevented me from working); WAI: Work Ability Index (score 0 = completely unable to work to 10 = work ability at its best); QQ: Quality and Quantity scales (score 0 = practically nothing/very poor quality to 10 = normal quantity/very good quality); QQ-Total: $Q_{quality}$; VAS: visual analog scale; HAQ: Health Assessment Questionnaire.

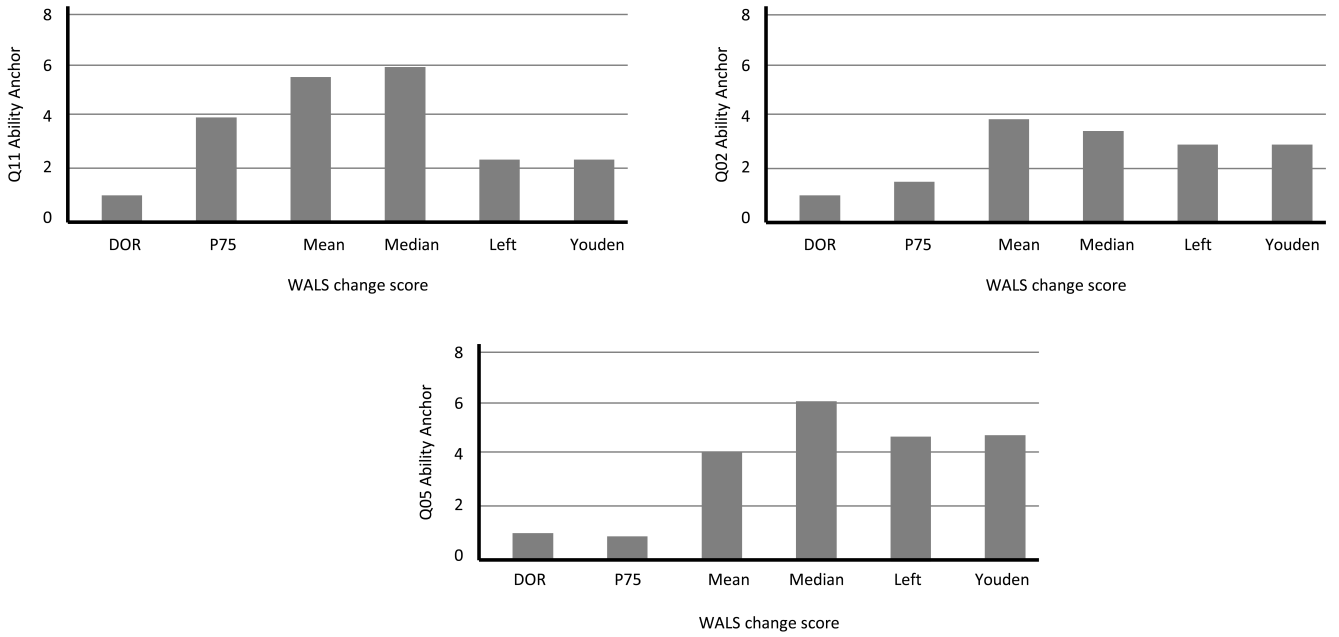

Figure 2. Variation in minimal important difference estimates for WALS score depending on anchors and analytical approaches. WALS: Workplace Activity Limitations Scale; DOR: Diagnostic Odds Ratio; P75: 75th percentile.

responding, but whether various MCID values would lead to different interpretations of the relative gains. Another approach discussed was the cumulative proportion responders analysis graph[20], which plots proportion responders (as defined by having exceeded the MCID) against magnitude of change with 1 line for each arm in a trial. For clinical trials, this allows more transparent interpretation of the difference between arms. In MCID development work, a plot for each MCID value in a cohort would allow us to see whether different MCID thresholds had a large or small difference in the proportion classified as improved. The breakout groups agreed that these reporting approaches could improve the management of multiple MCID values. They will be forwarded to the Technical Advisory Group of OMERACT for consideration.

Key points resulting from the SIG:

- A dual scale, measuring both time and difficulty, better identifies patient's experience, but the main drawback is the length and complexity of such a scale.
- There is no perfect global scale, but a generic scale with a recall > 1 day and < 1 month is preferred.
- Development of reporting approaches is key to improve management of multiple MCID values.

DISCUSSION

We have continued to gather the evidence needed to recommend the right worker productivity outcome measures to be included in clinical studies. Moving toward OMERACT 17:

- We are updating our literature against Filter 2.1 and will be finalizing our analysis of global scales for voting at OMERACT 17.
- We will further evaluate the value of the dual scale and test in other trials with an aim to recommend a better scale identifying both difficulty and time having difficulties.
- We will provide recommendations for PAS/MCID to be applied in worker productivity studies and to inform future MID/PAS research in other areas.
- We will further our understanding of contextual factors in relation to worker productivity loss and our work will inform the OMERACT contextual factor group.

ACKNOWLEDGMENT

We acknowledge representatives from BMS, AbbVie, UCB, and Pfizer for their collaboration with the OMERACT worker productivity group. We also acknowledge all researchers involved in the EULAR-PRO at-work productivity group for their contribution to the global measure studies.

REFERENCES

1. Tang K, Escorpizo R, Beaton DE, Bombardier C, Lacaille D, Zhang W, et al. Measuring the impact of arthritis on worker productivity: perspectives, methodologic issues, and contextual factors. J Rheumatol 2011;38:1776-90.
2. Beaton DE, Dyer S, Boonen A, Verstappen SM, Escorpizo R, Lacaille DV, et al. OMERACT filter evidence supporting the measurement of at-work productivity loss as an outcome measure in rheumatology research. J Rheumatol 2016;43:214-22.
3. Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745-53.
4. Boers M, Kirwan JR, Tugwell P, Beaton D, Bingham CO III, Conaghan PG, et al. The OMERACT Handbook. [Internet. Accessed April 4, 2019.] Available from: https://omeract.org/resources
5. Osterhaus JT, Purcaru O, Richard L. Discriminant validity, responsiveness and reliability of the rheumatoid arthritis-specific Work Productivity Survey (WPS-RA). Arthritis Res Ther 2009;11:R73.
6. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. Pharmacoeconomics 1993;4:353-65.
7. Tuomi K, Ilmarinen J, Jakhola A, Katajrinne L, Tulkki A. Work ability index. Helsinki: Finnish Institute of Occupational Health; 1998.

8. Brouwer WB, Koopmanschap MA, Rutten FF. Productivity losses without absence: measurement validation and empirical evidence. Health Policy 1999;48:13-27.
9. Gignac MA, Badley EM, Lacaille D, Cott CC, Adam P, Anis AH. Managing arthritis and employment: making arthritis-related work changes as a means of adaptation. Arthritis Rheum 2004;51:909-16.
10. Lerner D, Amick BC III, Rogers WH, Malspeis S, Bungay K, Cynn D. The Work Limitations Questionnaire. Med Care 2001;39:72-85.
11. Leggett S, van der Zee-Neuen A, Boonen A, Beaton DE, Bojinca M, Bosworth A, et al; At-work Productivity Global Measure Working Group. Test-retest reliability and correlations of 5 global measures addressing at-work productivity loss in patients with rheumatic diseases. J Rheumatol 2016;43:433-9.
12. Leggett S, van der Zee-Neuen A, Boonen A, Beaton D, Bojinca M, Bosworth A, et al; at-work productivity global measure working group. Content validity of global measures for at-work productivity in patients with rheumatic diseases: an international qualitative study. Rheumatology 2016;55:1364-73.
13. Leggett S, Boonen A, Lacaille D, Talli S, Bojinca M, Karlson, et al ; on behalf of EULAR-PRO at-work productivity co-investigators. Moderate to good construct validity of global presenteeism measures with multi-item presenteeism measures and patient reported health outcomes: EULAR-PRO worker productivity study [abstract]. Ann Rheum Dis 2017;76 Suppl 2:467-8.
14. Carruthers EC, Rogers P, Backman CL, Goldsmith CH, Gignac MA, Marra C, et al. "Employment and arthritis: making it work" a randomized controlled trial evaluating an online program to help people with inflammatory arthritis maintain employment (study protocol). BMC Med Inform Decis Mak 2014;14:59.
15. Tran K, Li XY, Seah XC, Backman C, vanAs B, Rogers P, et al. Process evaluation of the making it work program, an online program to help people with inflammatory arthritis remain employed [abstract]. Arthritis Rheumatol 2017;69 Suppl 10:197.
16. Kobza A, Beaton D, Gignac M, Lacaille D. Psychometric evaluation of a modified measure of presenteeism in inflammatory arthritis [abstract]. J Rheumatol 2017;44:940.
17. Donaldson M, Kobza A, Beaton DE, Gignac MA, Lacaille D. Measurement properties of presenteeism measures with dual answer keys in inflammatory arthritis [abstract]. Ann Rheum Dis 2017;76 Suppl 2:471.
18. Copay AG, Eyberg B, Chung AS, Zurcher KS, Chutkan N, Spangehl MJ. Minimum clinically important difference: current trends in the orthopaedic literature, part II: lower extremity: a systematic review. JBJS Rev 2018;6:e2.
19. McLeod LD, Coon CD, Martin SA, Fehnel SE, Hays RD. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. Expert Rev Pharmacoecon Outcomes Res 2011;11:163-9.
20. Farrar JT, Dworkin RH, Max MB. Use of the cumulative proportion of responders analysis graph to present pain data over a range of cut-off points: making clinical trial data more understandable. J Pain Symptom Manage 2006;31:369-77.