Considerations for evaluating and recommending worker productivity outcome measures: An update from the OMERACT Worker Productivity Group

S.M.M. Verstappen[1,2,3], D Lacaille[4], A Boonen[5], R Escorpizo[6,7], C Hofstetter[8], A Bosworth[9], A Leong[10], S Leggett[1], M.A.M. Gignac[11], J.K. Wallman[12], M.M. Terwee[13], F Berghea[14], M Agaliotis[15], P Tugwell[16], D Beaton[11].

[1] 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.
[2] NIHR Manchester Biomedical Research Centre, Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, UK
[3] Arthritis Research UK/MRC Centre for Musculoskeletal Health and Work, University of Southampton, UK.
[4] Department of Medicine, and the Division of Rheumatology, Department of Medicine, University of British Columbia, Vancouver; Arthritis Research Canada, Richmond, British Columbia
[5] Division of Rheumatology, Maastricht University Medical Center, Care and Public Health Research Institute (CAPHRI),, Maastricht, the Netherlands.
[6] Department of Rehabilitation and Movement Science, University of Vermont, Burlington, VT, USA.
[7] Swiss Paraplegic Research, Nottwil, Switzerland
[8] OMERACT Patient Research Partner, Canada.
[9] National Rheumatoid Arthritis Society (NRAS), Maidenhead, UK.
[10] OMERACT Patient Research Partner, Santa Barbara, CA, USA.
[11] Institute for Work & Health, Toronto, Ontario, Canada; University of Toronto, Toronto, Ontario, Canada
[12] Lund University, Skåne University Hospital, Department of Clinical Sciences Lund, Rheumatology, Lund, Sweden
[13] Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands.
[14] Carol Davila University of Medicine, Bucharest, Romania
[15] School of Public Health & Community Medicine, University of New South Wales, Australia
[16] 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.

Key words: OMERACT, presenteeism, psychometric properties, dual presenteeism scale, patient acceptable state, minimum important difference

Word count: 1668

Correspondence to: Suzanne 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

Abstract

Objectives. The OMERACT Worker Productivity group continues efforts to assess psychometric properties of measures of presenteeism.

Methods. Psychometric properties of single-item and dual answer multi-item scales were assessed plus methods to assess 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.

Background

Quantifying restrictions in worker participation, including absenteeism, sick leave, and presenteeism (i.e. reduced productivity due to 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 (RMDs) can experience variable levels of presenteeism and absenteeism dependin on their health status, job demands or other personal or environmental contextual factors (1).

During the last eight 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 (1-4). Based on a review of available instruments in the literature we had a mandate to move forward with six candidate measures (four single-item global and two multi-item 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 organised into a taxonomy of 4 different types of worker productivity measures, which sit against the background of contextual factors (Figure 1). At OMERACT12 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 OMERACT12 we have progressed in our research across the following four work-streams: i) collating further evidence about reliability, content/construct validity of global (i.e. single-item) measures of presenteeism and supplementing information on WPS-RA and WPAI which were previously endorsed; ii) evaluation of psychometric properties of dual answer scales of two validated multi-item measures; and iii) determination of patient acceptable state (PAS) and the minimal important difference (MID) of presenteeism measures; iv) contextual factors.

Materials and Methods

SIG OMERACT 2018.
At OMERACT14 we presented an update of our work on the first three work-streams. Attendees at our SIG included patients (n=4), clinicians (n=7), one fellow and others (e.g. methodologists, industry, n=5). Important questions were discussed with participants during break-out sessions, including:

- *Global measures: Based on the results presented (reliability, cross-cultural differences, construct validity) what would be your preferred global measure and why?*
- *Multi-Item 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?*
- *PAS / MID: i) How best to manage MID thresholds and ii) 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-Working trial: University of British Columbia Research Ethics Board (H11-03527); the EULAR-PRO study obtained overall ethical approval from NRES Committee NW–Greater Manchester (12/NW/0172) and from each participating centre 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 previously and show fair to excellent test-retest reliability (Intra Class Correlation (ICC) for 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 five measures, especially with respect to '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 impact 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 one 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). Furthermore, participants highlighted the difficulty in answering and interpreting disease specific scales, due to the complexity of many rheumatic diseases, and preferred a generic scale.

Multi-item measures

How to best measure presenteeism using multi-item 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 from a Canadian RCT (Making-It-Work Program) of an employment intervention including patients with IA were used (N=364) (14, 15). The psychometric properties of the measures were first evaluated with the two answer keys analysed separately (i.e. without combining results) (16). Answers from the dual answer keys were then combined into a single score, obtained by: i) multiplying or ii) 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 only found in 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 analysed separately and combined (except WLQ-Physical Demands)(16). As a priori hypothesized, moderate correlation were observed between original answer keys, or combined scores, of WLQ subscales and WALS with measures assessing similar concepts [WPAI, or work instability scale (WIS)] (congruent validity). During the SIG all agreed that dual answer keys provided additional value. Patient representatives uniformly 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 (felt to be difficult to answer by patients); and concern about % 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. patient acceptable state of pain (PAS)) 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 has demonstrated that there are considerable differences in MCID thresholds depending on the anchor or method (18). At OMERACT16 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 which becomes a gold standard for threshold determination, and the methods used to determine the actual cut-off. 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" (see example 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's are highlighted with a vertical line on the same graph and demonstrate not only the proportion 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 one 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 if 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 SIG

- A dual scale, measuring both time and difficulty, better captures 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 more than one day and less than one month is preferred.
- Development of reporting approaches is key to improve management of multiple MCID values

Summary

We have continued to gather the evidence needed to recommend the right worker productivity outcome measures to be included in clinical studies. Moving towards OMERACT17:

- We are updating our literature against filter 2.1 and will be finalizing our analysis of global scales for voting at OMERACT17.
- We will further evaluate the value of the dual scale and test in other trials with an aim to recommend a better scale capturing 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.

Acknowledgements

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

Conflict of interest

The authors have no conflict of interest.

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 KJ, Tugwell P, Beaton D, Bingham CO III, Conaghan PG, et al. The OMERACT Handbook. [Internet Accessed May 17, 2017] 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, Boonen A, Beaton DE, Bojinca M, Bosworth A, et al. 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, Boonen A, Beaton D, Bojinca M, Bosworth A, et al. Content validity of global measures for at-work productivity in patients with rheumatic diseases: an international qualitative study. Rheumatology (Oxford) 2016;55:1364-73.
13. Leggett S, Boonen A, Lacaille D, Talli S, Bojinca M, Karlson, et al. Moderate to good construct validity of global presenteeism measures with multi-item presenteeism measures and pateint reported health outcomes: EULAR-PRO worker productivity study [Abstract]. Ann Rheum Dis 2017;76 (supppl2):467-468.
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 LX, Seah XC, Backman C, van As, B, Rogers P, Gignac M, Esdaile J, Thorne C, Li L, Lacaille D. Process Evaluation of the Making It Work Program, an Online Program to Help People with Inflammatory Arthritis Remain Employed [Abstract]. Arthritis Rheum; 2017. p. 230-1.
16. Kobza A BD, Gignac M, Lacaille D [Abstract]. Psychometric Evaluation of a Modified Measure of Presenteeism in Inflammatory Arthritis. J Rheumatol 2017:44;940.
17. Donaldson M KA, 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 reviews. 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.

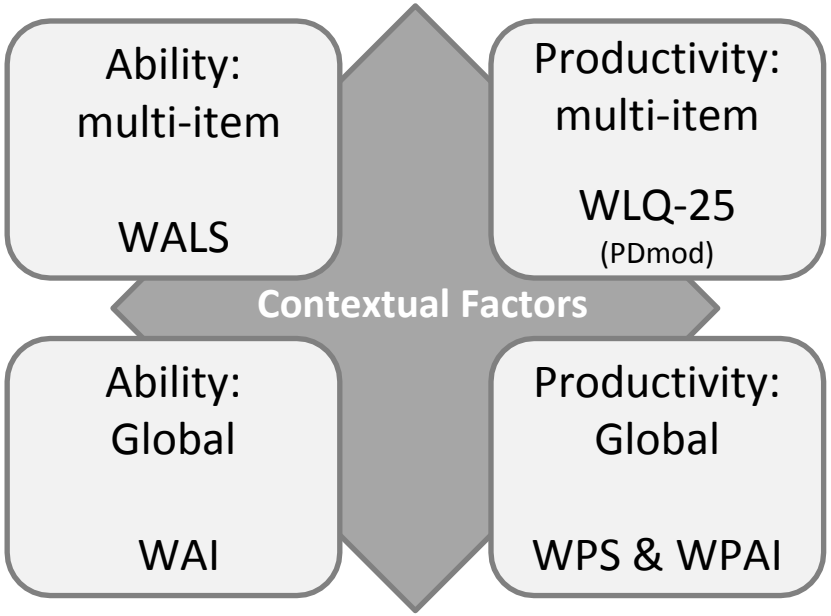

Figure 1. Organization into a taxonomy of 4 different types of work productivity measures. WALS: Workplace Activity Limitations Scale; WLQ-25 Pdmod: Work Limitations Questionnaire with modified physical demands scale; WAI: Work Ability Index; WPS: Arthritis-specific Work Productivity Survey; WPAI: Work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

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

		WPAI *r*	WPS-A *r*	WAI *r*	QQ-Quantity *r*	QQ-Quality *r*	QQ-Total *r*
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
Multi-item 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

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

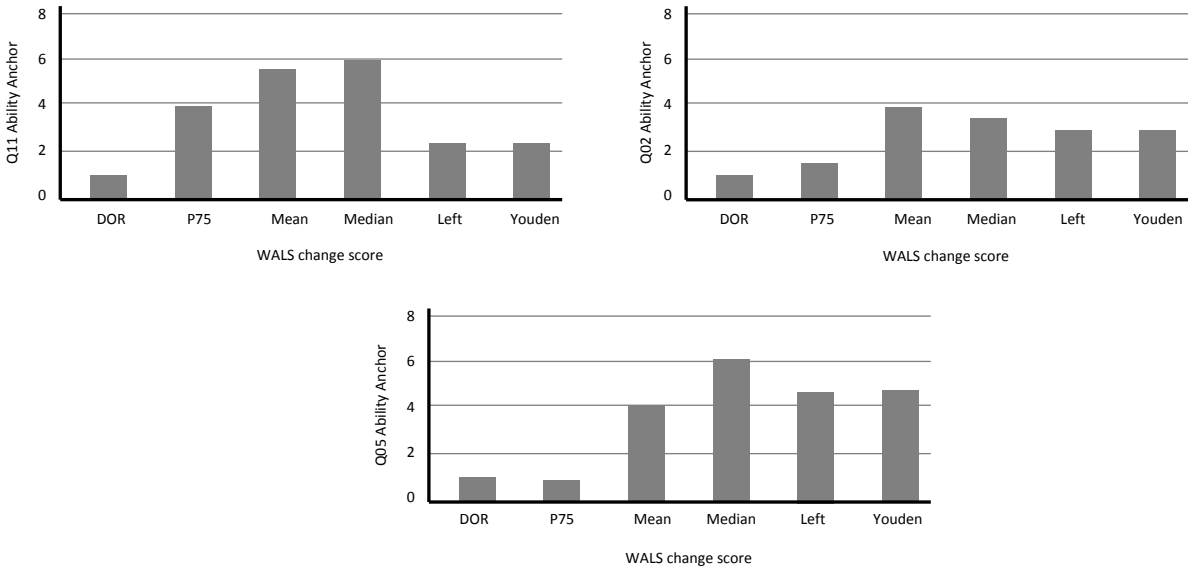

Figure 2. Variation in minimal important difference estimates for WALS score depending on anchors and analytical approaches. WALS = Workplace Activity Limitations Scale; P75 = 75th percentile.