Master's Thesis in Physiotherapy

May 2020

Measurement properties of the Norwegian version of the Physical Workload Questionnaire (PWQ)

Lise Grethe Kjønø

MAFYS5900

Word count thesis: 9146

Faculty of Health Sciences

OSLO METROPOLITAN UNIVERSITY
STORBYUNIVERSITETET

FOREWORD

I would like to use this opportunity to thank the ones contributing to realisation of this thesis. I want to express my gratitude to my supervisor Kjersti Storheim for useful comments, remarks and for sharing your knowledge and valuable time in excellent guidance through the learning process of this master thesis. Furthermore, I would like to thank my co-supervisor Ørjan Nesse Vigdal for useful help and comments on the way. I was fortunate enough to be given the chance to write my master thesis as part of the Norwegian BACE project, so thank you Margreth Grotle, project manager of BACE-N, for including me and providing me with data. Thanks also to Rikke Munk for your help and support in the process.

I would also like to extend my deepest gratitude to my loved ones, my dear Anna, Aksel and Tom Erik, who have supported me throughout the entire process. Thank you for letting me do this, it has been extremely important to have your support and for being so patience.

The past two years have been filled with challenges and hard work. However, I am very grateful for having had the opportunity to do this. I have learned a lot, I am happy and proud of what I have accomplished, but most of all I have proven that it is never too late to start something new.

Hagan,

May 2020

Lise Grethe Kjønø

SAMMENDRAG

Bakgrunn: «Spørreskjema for fysisk arbeidsbelastning» er et spørreskjema basert på 26 spørsmål som er relatert til fysisk arbeidsbelastning. Spørreskjemaet har nylig blitt oversatt til norsk, men skjemaets måleegenskaper har ikke blitt undersøkt. Hensikten med denne studien er å undersøke måleegenskapene til den norske versjonen av «Spørreskjema for fysisk arbeidsbelastning» hos pasienter med ulike muskelskjelettplager.

Metode: Et tverrsnitts studie og test-retest design ble benyttet for å undersøke spørreskjemaets indre struktur, interne konsistens, begrepsvaliditet og reliabilitet. Eksplorerende faktoranalyse ble brukt for å undersøke indre struktur og hvilke spørsmål som skulle inkluderes i den norske versjonen av skjema. Intern konsistens ble vurdert ved hjelp av Cronbach's alpha, og hypotesetesting ble brukt for å undersøke begrepsvaliditet («kjent»-gruppe, konvergent og diskriminerende validitet). Reliabilitet ble vurdert ved hjelp av korrelasjonskoeffisient ($ICC_{2.1}$), standard målefeil ($SEM_{agreement}$) og minste oppdagbare endring ($SDC_{95\%ind}$).

Resultater: 115 pasienter med muskelskjelettplager og med gjennomsnittsalder (SD) 46 (9) år ble inkludert i tverrsnittstudiet, hvorav 48 ble inkludert i test-retest analysene. Eksplorerende faktoranalyse av de 26 spørsmålene resulterte i to subskalaer: «Tung fysisk arbeidsbelastning» (15 spørsmål, score 0-100) og «Langvarige stillinger og repeterende bevegelser» (7 spørsmål, score 0-100). Intern konsistens viste en Cronbach's alpha verdi på henholdsvis 0.94 og 0.85 på subskala 1 og 2. Det var ingen gulv eller takeffekt av subskalaene. Undersøkelse av begrepsvaliditet avdekket at 12 av 14 (85%) pre-definerte hypoteser ble bekreftet. Test-retest reliabilitet av spørreskjemaet viste en $ICC_{2.1}$ på 0.96 (95% KI 0.88, 0.98) og 0.92 (95% KI 0.81, 0.96), $SEM_{agreement}$ på 6.9 og 10.0 og $SDC_{95\%}$ind på 19.2 og 27.7 på henholdsvis subskala 1 og 2.

Konklusjon: Den norske versjonen av «Spørreskjema for fysisk arbeidsbelastning» viser tilfredsstillende intern konsistens, begrepsvaliditet og reliabilitet, og kan brukes ved måling av fysisk arbeidsbelastning hos pasienter med ulike muskelskjelettplager.

Nøkkelord: Spørreskjema for fysisk arbeidsbelastning, fysisk arbeidsbelastning, muskelskjelettplager, validitet, reliabilitet

Table of Contents

LIST OF TABLES

Appendix

LIST OF FIGURES

ABBREVIATIONS

BACE: BAck Complaints in Elders

CI: Confidence Interval

COSMIN: COnsensus-based Standards for the selection of health Measurement INstruments

DMQ: Dutch Musculoskeletal Questionnaire

EFA: Exploratory Factor Analysis

ICC: Intraclass Correlation Coefficient

iPCQ: iProductivity Cost Questionnaire

MEI: Mechanical Exposure Index

MIC: Minimal Important Change

NRS: Numeric Rating Scale

NSD: Norwegian Centre for Research Data

PCA: Principal Component Analysis

PROMs: Patient-Related Outcome-Measures

PWQ: Physical Workload Questionnaire

QPSnordic: General Nordic Questionnaire for psychological and social factors at work

REC: Regional Ethical Committees for Medicine and Health

SD: Standard Deviation

SDC: Smallest Detectable Change

SEM: Standard Error of Measurement

SF-36: Short Form 36 Health Status Questionnaire

VDU: Visual Display Unit

1 INTRODUCTION

Physical work demands are thought to be associated with development of various musculoskeletal disorders (Da Costa & Vieira, 2010; Mayer, Kraus, & Ochsmann, 2012), and to be related to labour market participation (Andersen, Fallentin, Thorsen, & Holtermann, 2016; Sterud, 2014). Work-related musculoskeletal disorders are associated with sickness absence in a variety of occupations, and are among the leading causes of sickness absence in Europe (Bang, Lund, Labriola, Villadsen, & Bültmann, 2007; Griffith et al., 2012). According to existing literature, the major physical risk factors for work-related musculoskeletal disorders seems to be heavy lifting, working with a bent or twisted back or elevated arms, repetitive movements and vibration (Da Costa & Vieira, 2010; Mayer et al., 2012).

To prevent work-related musculoskeletal disorders, it is necessary to assess physical workload at the workplace (Stock, Fernandes, Delisle, & Vezina, 2005). Investigating the contribution of physical workload to musculoskeletal disorders may be based on direct measurements, observational methods or self-reported questionnaires (David, 2005). Direct measurements and observational methods are very resource demanding, especially in large epidemiological studies. Therefore self-reported questionnaires are commonly used and offer the possibility of studying a great number of persons at a modest cost, as well as allowing the investigation of a large number of variables when collecting exposure information (Stock et al., 2005). However, numerous studies have warned researchers about the lack of validity of self-reported questionnaires (David, 2005; van Der Beek & Frings-Dresen, 1998).

A wide variety of questionnaires have been developed to assess physical workload, but often their measurement properties have not properly been tested (Barrero, Katz, & Dennerlein, 2009; Stock et al., 2005). The Physical Workload Questionnaire (PWQ) was developed by Bot et al., with the aim to create a short and simple self-report questionnaire for assessing physical workload, to be used both in occupational health care as well as in epidemiological research (Bot et al., 2004). The items were captured from the Dutch Musculoskeletal Questionnaire (DMQ) which is a screening instrument for musculoskeletal workload and associated potential hazardous working conditions (Hildebrandt, Bongers, van Dijk, Kemper, & Dul, 2001). The items in the DMQ were based on reviews of the epidemiological literature. In contrast to other physical workload questionnaires, the DMQ has been studied thoroughly for its dimensionality and validity (Hildebrandt et al., 2001). However, the domain taken into account regarding musculoskeletal workload has 63 items and is therefore too lengthy to be used in most studies. The selection of the items for the PWQ were based on face validity and a discussion among experts during a consensus meeting. An item pool of 26 items that were expected to have an association with either upper or lower extremity complaints were chosen and tested for its dimensionality, internal consistency and construct validity in a population with upper extremity and lower extremity musculoskeletal disorders in the Netherlands. The item pool formed two subscales and the results supported the internal structure, internal consistency and construct validity (Bot et al., 2004).

To our knowledge the PWQ has not yet been translated to other languages or been tested for its measurement properties in other studies. Before a measurement instrument can be used in research or clinical practice, its measurement properties should be assessed and considered adequate (Mokkink et al., 2010a). The PWQ item pool is translated and cross-culturally adapted into Norwegian. To enable the use of the Norwegian version of the PWQ in clinical settings and research in Norwegian speaking patients, the present study aims to test its measurement properties in patients with musculoskeletal disorders.

1.1 Research question

Does the Norwegian version of the Physical Workload Questionnaire show acceptable internal consistency, construct validity and reliability for patients with various musculoskeletal disorders?

1.2 Project organization and anchoring

This study is conducted as part of the master programme in physiotherapy at Oslo Metropolitan University. It is part of the BACE (BAck Complaints in Elders) study in Norway which is a prospective cohort study designed to assess elderly (≥55 years) patients with back pain. BACE is an international consortium led by Chris Maher (Australia) and Bart Koes (The Netherlands). The Norwegian BACE study is led by professor Margreth Grotle at Oslo Metropolitan University. The PWQ is part of the comprehensive questionnaire used in the cohort study and has previously, according to international guidelines, been translated and cross-culturally adapted into Norwegian by researchers in the BACE project (appendix 2). The data used in this study has been collected previously for validation purposes of the PWQ.

The thesis is written as a scientific article, complemented by a more descriptive document (“kappe”). Supervisors were professor Kjersti Storheim and- PhD-student Ørjan Nesse Vigdal.

2 THEORETICAL FRAMEWORK

2.1 Physical workload

Physical workload refers to all external loads at workplace creating biomechanical forces to the limbs and trunk that further creates internal loads on the tissues and anatomical structures (National Research Council Institute of Medicine, 2001). Physical working conditions involves pushing and pulling loads, lifting, holding and carrying loads, whole-body forces, awkward body postures, dynamic body movements, repetitive manual work processes and mechanical exposures like vibrations (Tynes et al., 2017). Studies have revealed several risk factors in physical workload for development of musculoskeletal disorders. In a systematic review of longitudinal studies, Da Costa and Vieira (2010) found that the most commonly reported biomechanical risk factor for musculoskeletal disorders were excessive repetition, awkward postures and heavy lifting. In another systematic review, Mayer et al. (2012) found an association between development of shoulder and/or neck pain with manual material handling, repetitive work, vibration, trunk flexion or rotation and working with hands above shoulder height. Griffith et al. (2012) found in a meta-analysis an association between physical workload and low back pain, and Demarchi et al. (2019) found in another study an association between overload related to trunk postures and arm positions in patients with chronic low back pain. Based on existing literature; heavy lifting, working with a bent or twisted back or elevated arms, and using repetitive movements and vibrations seems to be the major physical risk factors (Da Costa & Vieira, 2010; Demarchi et al., 2019; Mayer et al., 2012; Sterud, 2014; Sterud & Tynes, 2013).

2.2 Measuring instruments to estimate physical workload

Measurement instruments developed for clinical practice or research can be used for various purposes: measurement of outcomes, discriminating between subjects, predicting either prognosis or the results of some other test, assessment of diagnosis and evaluating change over time (de Vet, Terwee, Mokkink, & Knol, 2011; Kirshner & Guyatt, 1985). To prevent work-related musculoskeletal disorders, it is necessary to assess physical workload at the workplace (Stock et al., 2005). There are several available methods for this, and the appropriate assessment method should be selected according to the study's aims, the applicability and validity of these methods, and economic aspects (David, 2005).

The assessment of physical workload may be based on direct measurement, observational methods or self-report questionnaires (David, 2005). A variety of direct technical assessment instruments such as accelerometery, pedometry, heart rate monitoring and indirect calorimetry are for use in the field (Tremblay, Colley, Saunders, Healy, & Owen, 2010). These instruments are believed both to be valid and associated with minor error in use, but are very resource demanding, especially in large studies (Stock et al., 2005).

There are several valid observational methods available to assess the physical workload. However, there is no gold standard and observational methods need skilled observers and consume much time, which has implications for costs and feasibility (Takala et al., 2010). Limiting the number and duration of observations may also reduce validity (Stock et al., 2005).

By using questionnaires, a wider period of time can be studied, all tasks of the job can be covered and physical workloads that more accurately reflect the general workload can be estimated (Stock et al., 2005). Although it is not possible to quantify the workload and only crude estimations of the amplitude, frequency, or duration of the workload can be made, information collected by questionnaire may be sufficient to rank the physical workload of specific activities, tasks or jobs (Burdorf & Van Der Beek, 1999). Additionally, questionnaires are more efficient and cheaper to use than direct measurement and observations, and can be used as screening methods and in larger studies (Dale, Strickland, Gardner, Symanzik, & Evanoff, 2010; Stock et al., 2005). However, questionnaires can be less accurate and reliable than direct measurement and observational methods. Awareness of the body posture and movements is difficult and may lead to over or underestimating the physical demands. The questions can also be interpreted differently to how it was intended by the developers (Stock et al., 2005). Self-reported questionnaires used to assess physical workload have shown varying validity, and are often tested against observational methods with their own strengths and limitations (Barrero et al., 2009; Stock et al., 2005; Takala et al., 2010).

2.3 Self-reported questionnaires to assess physical workload

A wide variety of questionnaires have been developed to assess physical workload, but often their measuring properties have not been properly tested (Barrero et al., 2009; Stock et al., 2005). Most of the questionnaires are composed of various items relating to physical load. Common items being assessed are general body posture (e.g. sitting, standing, walking), posture of the neck, shoulders, arms, wrists or hands, repetitive movements, hand use, vibration, and level of overall physical effort (Stock et al., 2005).

A questionnaire developed by Hollmann, Klimmer, Schmidt, and Kylian (1999) is based on a biomechanical model where the items are summed, with the intention to describe forces in the lumbar spine which gives an estimated physical workload (Hollmann et al., 1999). In other questionnaires the items are analysed separately (Leijon, Wiktorin, Härenstam, & Karlqvist, 2002; Viikarijuntura et al., 1996) whereas some questionnaires have been divided into one (Balogh et al., 2001) or several subscales (Pope, Silman, Cherry, Pritchard, & Macfarlane, 1998; Wiktorin, Hjelm, Winkel, & Köster, 1996).

The Dutch musculoskeletal questionnaire (DMQ), developed by Hildebrandt et al. (2001), is a screening instrument for the analysis of musculoskeletal workload and associated potential hazardous working conditions. The DMQ consists of nine pages with approximately 25 questions per page and the domain taken into account about musculoskeletal workload has 63 items expressed in questions about postures, forces and movements. In contrast to other physical workload questionnaires, this one has been studied thoroughly for its dimensionality and validity. Convergent and divergent validity were found to be fair and there were evidence for concurrent validity (Hildebrandt et al., 2001). Due to its number of items it is too lengthy to be used in large epidemiological studies.

2.4 The Physical Workload Questionnaire

The PWQ is a self-report questionnaire for assessing physical workload in occupational healthcare and in research (Bot et al., 2004). The questionnaire was based on an item pool consisting of 26 items assessing force, dynamic and static load, repetitive load, (uncomfortable) postures, sitting, standing and walking. In the only former study, the item pool was assessed for its dimensionality, internal consistency and construct validity in a population with upper extremity and lower extremity musculoskeletal disorders. Factor analysis revealed two subscales; twelve items related to the first subscale “heavy physical work” and six items related to the second subscale “long lasting postures and repetitive movements”. The remaining eight items (“prolonged sitting”, “prolonged visual display units work”, “work with vibrating tools”, “operate peddles with feet”, “climbing stairs”, “twisted posture”, “uncomfortable postures” and “walking on irregular surfaces”) were excluded due to low loadings or to similar loadings on both factors. Both subscales showed good internal consistency evaluated by Cronbach’s alpha (Bot et al., 2004). The construct validity was examined based on hypotheses that physical workload would vary among different occupational groups, and Bot et al. (2004) classified the occupations of the participants in their study into four groups on the basis of expected workload; “heavy physical load”, “long lasting postures and repetitive movements”, both heavy physical load and long lasting postures and repetitive movements” and “no physical load”. Six of eight hypotheses regarding the construct validity of the subscales were confirmed, and the validity of the questionnaire developed by Bot et. al. was considered to be good. Each item is scored on a 4-point Likert scale with the response options; “seldom or never” (0), “sometimes” (1), “often” (2), and “(almost) always” (3). Scoring is done by adding up the response to each item, which produces a raw score. The final scores are calculated by dividing the raw score by the maximum score possible on the subscale, multiplied by 100, resulting in a final score ranging between 0 (no workload) and 100 (highest workload) for each subscale (Bot et al., 2004).

2.5 Methods for evaluating the quality of an assessment tool

Before a measurement instrument can be used in research or clinical practice, its measurement properties should be assessed and considered adequate (Mokkink et al., 2010a). To provide high methodological quality of measurement instruments in a study, standards and criteria needs to be followed. Quality criteria for measurement properties of health status questionnaires were created by Terwee et al. (2007). Also, a group of experts; the COSMIN-group (The COnsensus-based Standards for the selection of health Measurement Instruments), standardized terminology and definitions of measurement properties and developed a taxonomy of measurement properties relevant for evaluating health instruments. This consensus on taxonomy, terminology and definitions is a tool to better evaluate measurement properties in health status questionnaires (Mokkink et al., 2010b). Figure 1 shows the taxonomy of relationships of measurement properties designed by the COSMIN panel.

In methodological research the goal is to document and improve qualities of clinical and research instruments (Carter & Lubinsky, 2015). Data of interest in methodological research is referred to as measurement properties, of which reliability and validity are a part of (de Vet et al., 2011). Reliability and validity indicate if a measurement is trustworthy, which means both to be consistent and to measure what it is supposed to measure, and should always be evaluated before using a measurement (Carter & Lubinsky, 2015).

2.5.1 Reliability

Reliability is an essential requirement for the measurements used in clinical practise and research, and it is a prerequisite for validity (de Vet et al., 2011). Reliability refers to the degree to which the measurement is free from measurement error and can be subdivided into three measurement properties: internal consistency, reliability and measurement error (Mokkink et al., 2010a).

Internal consistency is a measure of the degree of interrelatedness among the items in a scale or subscale (Mokkink et al., 2010a). It measures the extent to which items in a questionnaire assess the same construct, and it is an important measurement property in questionnaires intending to measure a single underlying construct by using multiple items (de Vet et al., 2011). Cronbach's alpha should be used to calculate each scale or subscale separately. A low Cronbach's alpha indicates a lack of correlation between the items in a scale, and opposite a high Cronbach's alpha indicates a high correlation between items within a scale (Terwee et al., 2007). The item-total correlation are the correlations between each item and the total score from the questionnaire which gives an indication whether or not the item is part of the questionnaire (de Vet et al., 2011) . The items should correlate with the total, and if values are less than 0.3, the item does not correlate well with the scale overall and should be considered excluded (Pallant, 2016).

The measurement property reliability refers to the extent to which individuals' scores who have not changed are the same for repeated measurement under several conditions: test-retest (over time), interrater (different persons on the same occasion) or intrarater (same persons on different occasions) (Mokkink et al., 2010b). In test-retest measurement of a questionnaire, the assessment implies that responses to the questionnaire items remain relatively consistent across repeated administration of the same questionnaire (Terwee et al., 2007). The time period between the tests should be long enough to prevent recall, though short enough to ensure that no change has occurred, normally 1 or 2 weeks are appropriate (Terwee et al., 2007).

Choosing the appropriate statistical method depends on the nature of the data. The ICC is the most suitable reliability parameter for continuous measures, while the weighted Cohens Kappa should be used for ordinal measures (Terwee et al., 2007). Different sub forms of ICC are available for different study designs, and the most commonly used are $ICC_{1.1}$, $ICC_{2.1}$ and $ICC_{3.1}$ (de Vet et al., 2011). ICC 0.70 is recommended as a minimum standard for reliability (Nunnally & Bernstein, 1994). Another way to express reliability is to assess the measurement error, which is the difference between a measured score and its true value, including both random errors and systematic error (de Vet et al., 2011). The measurement error can be expressed as the standard error of measurement (SEM), either including systematic differences ($SEM_{agreement}$) or excluding them ($SEM_{consistency}$) (Terwee et al., 2007). The measurement error expressed in SEM can according to de Vet et al. (2011) be measured in three ways. Many researchers are using the formula SEM = $SD\sqrt{1-ICC}$. Another way to derive the SEM value is from the error variance in the ICC formula, although this may not include the systematic error. To include the systematic error in the SEM, the VARCOMP analysis is recommended. This is done by taking the difference of the values of the two measurements, and calculate the mean and the SD of these differences: $SEM_{agreement}(o^2_o+o^2_{po, e})$, where o^2_o is the variance due to systematic error between observations and $o^2_{po, e}$ is the random error (de Vet et al., 2011).

Figure 1. The taxonomy of measurement properties designed by the COSMIN panel (Mokkink et al., 2010b). Reprinted with permission from the COSMIN group.

2.5.2 Validity

The degree to which an instrument truly measures what it is supposed to measure is called validity (Mokkink et al., 2010b). There are three different types of validity; content validity, criterion validity and construct validity.

Content validity shows to what extent the content of the instrument is representative for the construct one intends to measure (Mokkink et al., 2010b). When using multi-item questionnaires, all of the items should be relevant and including, and cover all aspects of the construct to be measured (de Vet et al., 2011).

Criterion validity is applicable in studies where there is a gold standard for the construct to be measured, and refers to the extent to which scores on a measure are correlated to the gold standard (de Vet et al., 2011). If there is a lack of a gold standard, construct validation should be used to provide evidence of validity (de Vet et al., 2011).

Construct validity refers to the degree to which the instrument measures the construct it purports to measure and includes structural validity, hypotheses testing and cross-cultural validity (Mokkink et al., 2010b). Structural validity of a questionnaire can be assessed using factor analysis, exploratory or confirming (Terwee et al., 2007). Factor analysis is a technique that is used to reduce a large number of variables/items into fewer numbers of factors (Pallant, 2016). Exploratory factor analysis (EFA) is common to use when there is no prior theory about the number of dimensions and the aim is to explore and identify the underlying structure of the variables (Pallant, 2016). Within EFA, principal components analysis (PCA) and common factor analysis (FA) can be distinguished. Both techniques attempt to produce a smaller number of linear combinations of the original variables in a way that captures most of the variability in the pattern of correlations (Pallant, 2016). In PCA the original variables are transformed into a fewer numbers of factors, with all of the variance in the variables being used (de Vet et al., 2011). In FA, factors are estimated using a mathematical model, whereby only the shared variance is analysed. Despite the theoretical principles of PCA and common FA differ, the results are often similar. PCA is the simplest method and most often used in studies (de Vet et al., 2011).

Hypotheses testing refers to the degree to which the scores of an instrument relates to other measures in a manner that is consistent with theoretically derived hypotheses based on the assumption that the instrument validly measures the construct to be measured (Mokkink et al., 2010b). Predefined hypotheses which need to be as specific as possible should be assessed to test for construct validity (Terwee et al., 2007). Hypotheses can be about expected differences in scores between “known” groups, convergent validity assesses how well a questionnaire is correlated with other questionnaires that measure the same construct or qualities, while discriminant validity tests whether instruments that are not supposed to be related are actual unrelated (de Vet et al., 2011). According to the COSMIN group, the construct validity is good if 75% of the predefined hypotheses are confirmed (Terwee et al., 2007). Cross-cultural validity refers to translating and culturally adapting measurements into another language (de Vet et al., 2011).

Responsiveness to change is a measurement property related to validity and reliability (de Vet et al., 2011). If there is poor reliability in an instrument, the patient must report a large improvement for a change to be visible. An instrument with poor validity (i.e. does not measure the construct intended), will have low responsiveness because the construct that has improved is not the one being measured (de Vet et al., 2011).

Floor and ceiling effects can occur when a proportion of individuals achieve the highest or lowest possible score of a scale and are considered present when more than 15% of the individuals achieve these values (Terwee et al., 2007). If many responders have the same lowest or highest score, they cannot be distinguished from each other, thus reliability is reduced (Terwee et al., 2007).

3 METHODS

The Consensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist was applied in the design of the master study (Mokkink et al., 2010a), in addition the PWQ will be assessed according to guidelines for Patient-Related Outcome-Measures (PROMs) (Terwee et al., 2007). Methodological details are outlined in the article.

3.1 Design

This is a quantitative study where the Norwegian version of the PWQ was tested for its measurement properties by using a cross-sectional design. In addition, a test-retest assessment was conducted at the patients second attendance to the clinic.

3.2 Participants

Participants were recruited from an outpatient rehabilitation clinic. Eligible participants were patients with musculoskeletal disorders, aged 18 or above, working or on sick leave. Exclusion criteria were patients being unable to speak, read or write in Norwegian.

3.3 Procedures and measurement

At baseline (the cross-sectional study) the included participants completed the PWQ items as part of a comprehensive questionnaire, which also included sociodemographic variables, pain localization, intensity and history, psychosocial work environment, health-related productivity costs and health-related quality of life. In addition, another questionnaire was used to assess physical workload. Patients consenting to participate at the reproducibility part of the study filled out the PWQ at their second attendance. Additionally, a global question recording change in the work condition in the time interval was completed.

3.4 Analysis

All data analyses were performed using SPSS version 26 (SPSS inc., Chicago, IL). Demographic data were described using descriptive statistics. Numerical variables were expressed as means and standard deviation (SD) for normally distributed variables while median and minimum-maximum values were used for non-normally distributed variables. Categorical variables were expressed as frequencies.

3.4.1 Exploratory factor analysis

EFA was used to determine the factorial structure of the PWQ. The suitability of data for factor analysis was confirmed using the Kaiser-Meyer-Olkin measure of sampling adequacy, a significant Bartlett`s Test of Sphericity and inspection of the correlation matrix (Tabachnick & Fidell, 2014).

PCA was used to extract the factors of the 26 items from the item pool. The number of factors to be retained was guided by three decision rules: Kaiser´s criterion, retention of eigenvalues above 1, Cattel´s scree plot (Cattell, 1966), and by the use of Horn´s parallel analysis (Watkins, 2000). To aid in the interpretation of the retained factors, factor loadings after direct oblimin rotation was computed (Tabachnick & Fidell, 2014). Next step involved interpreting the rotated solution by identifying which items loaded on each retained factor. Items with factor loading below 0.5 (Nunnally & Bernstein, 1994), and communalities value below 0.3 were excluded (Pallant, 2016). Items which cross-loaded were retained in the factor it loaded most strongly.

3.4.2 Internal consistency

The internal consistency of the subscales was examined using Cronbach's alpha. Cronbach's alpha between 0.70 and 0.95 gave a positive rating (Terwee et al., 2007). The item-total correlation was examined and items with values below 0.3 was considered excluded (de Vet et al., 2011).

3.4.3 Data quality

Proportions of missing data and floor and/or ceiling effects were described. Floor or ceiling effects were considered to be present if more than 15% of the patients reported the lowest or the highest possible score, respectively (Terwee et al., 2007).

3.4.4 Construct validity

Construct validity was assessed by hypotheses testing: "known" group validity, convergent validity and discriminant validity. "Known" group validity was tested with the same procedure as in the original study of Bot et al. These eight hypotheses were rejected at significance levels of 1% ($p < 0.01$) (Bot et al., 2004). To assess convergent and discriminant validity, another six hypotheses based on other health-related constructs were predefined. Correlation coefficients under 0.3, between 0.3 and 0.6 and over 0.6 were considered low, moderate and high, respectively (Andresen, 2000). If more than 75% of the predefined hypotheses were confirmed, the construct validity was good (Terwee et al., 2007).

3.4.5 Reliability and measurement error

A paired t-test was used to assess the mean difference between test and retest. An intraclass correlation coefficient ($ICC_{2,1}$), two-way ANOVA random effect model for absolute agreement was used to assess relative reliability, and standard error of measurement ($SEM_{agreement}$) and smallest detectable change ($SDC_{95\%ind}$) were used to analyse absolute reliability (measurement error). ICC of ≥ 0.70 was considered acceptable (Terwee et al., 2007).

3.5 Ethics approval and consent to participate

The general ethical principles for medical research, stated in the Declaration of Helsinki, were followed. All participants received written study information and a written informed consent was signed prior to inclusion in the study for all participants (appendix 3). Participation was voluntary, and all participants could decide to leave the study at any time without giving a reason.

The study is considered a quality assessment project by the Regional Ethical Committees for Medicine and Health (REC) (reference no. 2014/1634/REK vest) and was approved by the Norwegian Centre for Research Data (NSD) (reference No. 42149) in 2019.

4 RESULTS

A summary of the results and some supplementary details not included in the article are presented here in “kappa”. A total of 115 patients with a mean (SD) age of 46 (9) were included in the cross-sectional study. The majority of the participants were women (68.7%), most of the included participants (90%) were in paid work and the working hours per week were median 37.5 hours (range 7.5-52 hours). On average they reported moderate pain, the majority had pain for more than 3 months, the most frequently reported pain area was the back region and physical workload were in general low. Sixty-two participated in the test-retest study of which 48 reported no change in working conditions from test to retest and had complete PWQ scores. Patients participating in the test (n=115) and the re-test (n=48) were similar, but individuals participating in the re-test had slightly different pain site locations and slightly different scores on physical function and general health on SF-36. The time interval between test and re-test was median 3 days (range 1-10 days).

4.1 Exploratory factor analysis

The interpretation of the analyses revealed 22 items to be remained in two factors. The first factor: subscale 1 “Heavy physical workload” consisted of 15 items and the second factor: subscale 2 “Long lasting postures and repetitive movements” consisted of 7 items.

4.2 Internal consistency

The Cronbach alpha value was 0.94 and 0.85 for subscale 1 and 2, respectively. The item-total correlation was 0.53-0.84 and 0.52-0.73 for subscale 1 and 2, respectively.

4.3 Data quality

There were no floor or ceiling effects of the subscales. Information about missing data and floor and ceiling effects for the subscales and items separately are presented in table 1.

4.4 Construct validity

For the “known” group validity, six of the eight hypotheses were statistically significant different ($p<0.01$). For the convergent and discriminating validity, all six hypotheses were confirmed. In total, 12 (85%) of the 14 predefined hypotheses were confirmed.

4.5 Reliability and measurement error

Reliability of PWQ data demonstrated an $ICC_{2.1}$ of 0.96 (95% CI 0.88, 0.98) and 0.92 (95% CI 0.81, 0.96), SEM of 6.9 and 10.0 and $SDC_{95\%}$ind of 19.2 and 27.7 of subscale 1 and 2, respectively.

Table 1 Missing data and floor- and ceiling effects for the PWQ items and subscales (n=115)

PWQ items. Does your work involve…	**Missing, n (%)**	**Lowest (%)**	**Highest (%)**
Heavy physical workload	**8 (7.0)**	**9.3**	**0.0**
1. Standing	1 (0.9)	2.2	20.2
4. Walking	1 (0.9)	42.1	8.8
5. Kneeling/ squatting	2 (1.7)	64.6	3.5
7. Twisted posture	1 (0.9)	48.2	5.3
11. Hands above shoulders	1 (0.9)	58.8	3.5
12. Hands below knees	1 (0.9)	67.5	0.0
13. Moving loads (>5 kg)	1 (0.9)	49.1	11.4
14. Moving loads (>25 kg)	2 (1.7)	63.7	0.9
15. Exert force with arms	2 (1.7)	42.5	14.2
16. Maximal force exertions	1 (0.9)	59.6	7.9
17. Physical hard work	2 (1.7)	59.3	6.2
20. Work with vibrating tools	1 (0.9)	79.8	2.6
23. Often squatting	2 (1.7)	52.2	6.2
24. Walking on irregular surfaces	1 (0.9)	68.4	4.4
25. Sitting/ moving on knees	4 (3.5)	72.2	1.8
Long lasting postures and repetitive movements	**5 (4.3)**	**1.8**	**4.5**
6. Repetitive movement	1 (0.9)	14.9	34.2
8. Neck bent forward	1 (0.9)	30.7	13.2
9. Turning/ bending neck	2 (1.7)	27.4	11.5
10. Wrists bent or twisted	2 (1.7)	29.2	16.8
18. static posture	1 (0.9)	14.9	38.6
19. Uncomfortable posture	3 (2.6)	32.1	8.9
26. Repetitive tasks arms/hands	1 (0.9)	29.8	34.2

5 DISCUSSION OF METHODS

5.1 Study design

This is a cross-sectional study including a test-retest design. Cross-sectional study design is a type of observational study and are used to document the status of a group at a particular point in time to assess the prevalence of outcome or exposure of risk factors (Carter & Lubinsky, 2015). These studies can usually be conducted relatively fast and are inexpensive in use. As this is a one-time measurement, it cannot take into consideration the dimension of time and therefore not assess causality (Carter & Lubinsky, 2015). However, in the present study the aim was to assess exposure of physical workload and to study its association to other constructs, therefore a cross- sectional study is suitable. To determine the reliability, test-retest design is common to use. Measurements are made twice on a set of individuals at some specific interval of time, and the variability of the measurements can be compared (Carter & Lubinsky, 2015).

5.2 Participants and reflections on validity

According to COSMIN recommendations there should be more than 100 participants in a validity study. A total of 115 participants were included in the present study, which is a strength of the study. However, selection of a representative sample from the population is important to provide validity, which is a central term in research as it considers how much we can trust the results: *internal validity* (Rothman, Greenland, & Lash, 2008). The internal validity is related to the ability of a study to handle any systematic error that may give incorrect estimates and thereby inaccurate associations. This can be threatened by many factors, including the selection of participants in the study, *selection bias*, and errors in measurement, *information bias* (Rothman et al., 2008). Once the internal validity of the study is established, we can make a judgement regarding how transferable these results are to other populations: *external validity*. External validity is concerned with whom, in what setting, and at what time the results of research can be generalized. Lack of internal validity implies that the results from the study deviates from the truth, and therefore we cannot draw any conclusions (Rothman et al., 2008).

5.2.1 Selection bias

Selection bias reflects any error in selection of the study participants or from factors affecting the study participation. A consequence of this is that the relationship between exposure and outcome differs between the participants included in the study and those potentially eligible for the study (Rothman et al., 2008). The theoretically eligible participants in the present study would be all Norwegians in work or on sick leave with musculoskeletal disorders, and ideally the sample should be picked using simple random sampling with every subject standing an equal chance of being included. Our sample are recruited from a clinic located in Asker, a wealthy town close to the capital of Norway, which may imply that this study's population consist of participants with high socioeconomic status. Previous studies show that low socioeconomic status is associated with higher exposure of physical workload (Aittomäki, Lahelma, Roos, Leino-Arjas, & Martikainen, 2005; Mehlum, Kristensen, Kjuus, & Wergeland, 2008). Although the sample in this study covers a variety of occupations, the score on the "Heavy physical workload" subscale is relatively low and the classification of occupational groups resulted in few participants classified in the heavy physical load group. This may indicate that participation from employees with very demanding work was low, and there is reason to believe that by recruiting participants from a wider geographical area we would have reached a broader population regarding occupational variation.

Eligible participants were patients with musculoskeletal disorders. However, there were no information whether the patients had specific diagnosis or non- specific musculoskeletal pain, we can therefore not draw any conclusions in direction of specific diagnosis groups to generalise the results to. Furthermore, the exclusion criteria being unable to speak, read or write in Norwegian was important to ensure that the participants were able to understand and respond correctly to the questionnaire. However, we may have excluded participants with minority background which could have given a broader sample of the population by using these criteria. Although 14 of the participants were not in paid job (occupational rehabilitation, student, homemakers, disabled and unemployed), they were included in the analyses as they responded to the questionnaire, and we think that physical workload can occur regardless if the job is paid or not. However, these participants may have affected the results. There may also be bias due to subjects who declined to participate if these participants are different from those who were investigated. Due to limited resources it was not possible to record information on all patients attending the rehabilitation clinic during the data collection period. In addition, more women than men are represented in this study, which could threaten the generalisability of the results. However, musculoskeletal disorders are shown to be more prevalent in women (Ihlebæk, Brage, Natvig, & Bruusgaard, 2010), and also that women are more prevalent patients in outpatient clinics (Kinge, Knudsen, Skirbekk, & Vollset, 2015).

Therefore, the gender distribution of the present study is probably representative of the patients seeking treatment for musculoskeletal disorders in outpatient physiotherapy clinics.

The COSMIN group recommends a minimum of 50 participants in test-retest analyses. Although 62 participated in the test-retest study, we had to exclude 5 participants who reported change in the working condition from test to retest, and another 9 participants who had incomplete questionnaires. We could include one more participant in subscale 1 and two more in subscale 2 as they completed one of the subscales, and in that way meet the COSMIN requirements for subscale 2, and only be one too short on subscale 1. However, we chose to include only those who had complete questionnaires for both subscales, hence only 48 participants were included in the test- retest analyses. It is unlikely that one and two more responses would have changed our conclusions, but some imprecision in our estimates is possible due to low numbers of participants.

5.2.2 Information bias

Information bias is related to inaccurately measured information of the variables of interest. Self- reported questionnaires are frequently used to collect information in cross-sectional studies, however care must be taken in interpreting and drawing conclusions based on such information, as self-ratings may suffer from misclassification. Although self-reported questionnaires are known for being sufficient to rank the physical workload of specific activities, tasks or jobs (Burdorf & Van Der Beek, 1999), they have shown varying validity in several studies (Barrero et al., 2009). There are studies showing that workers with musculoskeletal disorders may overestimate the physical load compared to healthy workers (Balogh et al., 2004; Hansson et al., 2001), and even when participants are motivated to report the workload accurately, they may have difficulties with recalling and accurately reporting the information (Althubaiti, 2016). This may threaten the validity of the questionnaire.

5.3 The Physical Workload Questionnaire

Since the questions drawn from the DMQ, forming the basis for the development of PWQ only have been examined once previously (in the original study of Bot et al.), all 26 questions extracted from DMQ were translated into Norwegian and included in the factor analyses. A further argument to include all 26 questions was that the items in the item pool were originally chosen because they were expected to have an association with either upper- or lower extremity complaints. This is a major limitation of our study and actually raises the question of whether this questionnaire can be used to assess physical workload in any other population than patients with upper- or lower extremity complaints. However, the item pool consists of several items known in the literature as risk factors for development of different musculoskeletal disorders, including back pain; heavy lifting, working with a bent or twisted back, elevated arms, repetitive movements and vibrations (Da Costa & Vieira, 2010; Demarchi et al., 2019; Mayer et al., 2012; Sterud, 2014; Sterud & Tynes, 2013). In addition, the results from our factor analysis, which included four more items in the subscales that can be associated with for example back complaints, may indicate that the items do cover more aspects than only upper- and lower extremity complaints, and that the questionnaire can be used to assess physical workload among patients with various musculoskeletal disorders. However, we still have to be aware of that there might be dimensions not taken into account to cover all aspects of musculoskeletal disorders. This may be problematic in the use of the PWQ in the BACE study.

5.4 Exploratory factor analysis

EFA was done according to guidelines, and interpretation of the results were done in line with recommended threshold values (de Vet et al., 2011; Nunnally & Bernstein, 1994; Pallant, 2016; Tabachnick & Fidell, 2014). Our sample of 115 participants were considered satisfying to ensure stability of variance as it is recommended 4-10 subjects per item, with a minimum of 100 when performing FA (Kline, 1993). One challenge in the interpretation was whether to extract two or three factors. The scree plot didn't show a clear "elbow", indicating either a two- or three factor solution, whereas parallel analysis suggested two factors. Although parallel analysis is known for being a more accurate approach to estimating the number of factors, as both Kaiser`s criterion and the scree test is tending to overestimate the number of factors (Hubbard & Allen, 1987; Zwick & Velicer, 1986), we chose to examine both the two-factor and the three- factor solution. Therefore, the number of factors to extract was based on interpretation of the analyses of the two solutions. As a result, the two-factor solution was chosen due to one factor consisting of only two items in the three-factor solution. Hence, we assume that the EFA did not result in an under- or over factoring of the PWQ. The adequate results of the measurement properties of the PWQ support this conclusion.

5.5 Internal consistency

Cronbach's alpha is recommended to measure the internal consistency of the subscales (Terwee et al., 2007). An alpha exceeding 0.9, which is the case for the "heavy physical workload" subscale, may indicate that some items are redundant, which means they are testing the same question but in a different way (Nunnally & Bernstein, 1994). Examination of the item-total statistics shows that two items would decrease the Cronbach's alpha to 0.93 if they were removed from the scale (moving loads >5kg and physical hard work), however the gap is so minimal and we consider the items to be important and to contribute to the content validity of the instrument. The Cronbach's alpha coefficient can be sensitive to the number of items in a scale, and questionnaires or subscales with fewer than 10 items can result in a value that is too low (de Vet et al., 2011; Pallant, 2016). Despite this, the "long lasting postures and repetitive movements" subscale which consist of only 7 items, had a Cronbach's alpha of 0.85. Item-total correlation ranged from 0.53 to 0.84 in subscale 1, and from 0.52 to 0.73 in subscale 2, indicating all items to correlate well with the total subscales (de Vet et al., 2011).

5.6 Data quality

If more than 15% of the patients reported the lowest or the highest possible score, floor or ceiling effects were considered to be present (Terwee et al., 2007). Although there were no floor or ceiling effects of the subscales according to the definitions of Terwee et al. (2007), substantial proportions of participants had the lowest and highest possible score for several of the single items in the PWQ. These effects may be problematic because they weaken the ability to distinguish participants in the higher or lower levels from each other, thus reliability is reduced. If the questionnaire is used for measurement of change of workload, responsiveness may also be limited because changes cannot be measured (Terwee et al., 2007). In a group of patients with different occupations, it was expected to have some degree of floor or ceiling effect of some of the items separately. The fact that we found floor effect on almost all items, except from the items "repetitive movement" and "static posture" may be a call for a change in the response categories. This finding may also indicate that the population being studied is not well represented by the general population in terms of mean and distribution. On the other hand, as there were no floor or ceiling effect of neither of the subscales, this result may be interpreted as an indication that the PWQ items adequately covered the variation of physical workload in the different occupations.

5.7 Construct validity

Construct validity is an important element of the validity of a questionnaire. The structural validity is previously described. As there is no gold standard to relate the PWQ results with, construct validation by hypotheses testing is recommended to provide evidence of validity (de Vet et al., 2011). Predefined hypotheses which need to be as specific as possible should be assessed to test for construct validity (Terwee et al., 2007). We chose to use the same predefined hypotheses as Bot et al. (2004) did in the original study, "known" group hypotheses, where the occupations were classified into four different groups based on their expected workloads and then used as a gold standard. In line with Bot et al. (2004), these hypotheses were rejected at significance levels of 1% ($p < 0.01$). By doing this we could compare our results with the original study. However, this classification method is not standardized and as there is a large interindividual variability within jobs, there is a risk of misclassification of a substantial number of subjects implied. In the process of the classification of the occupational groups, several disagreements were whether to classify jobs in the heavy physical load group or in the group with both heavy physical load and long lasting postures and repetitive movements. It is possible that occupations classified in the heavy physical load group also involved long lasting postures and repetitive movements, and therefore these would score high on the "long lasting postures and repetitive movements" subscale as well. As the classification of the occupations resulted in relatively low number of participants in the heavy physical load group (n=10) it may imply that there would be a better solution to merge the heavy physical load group with the group classified as both heavy physical load and long lasting postures and repetitive movements, both to have a group with more participants and to avoid the misclassification issue with these two groups. On the other hand, this would make it difficult to distinguish between the two subscales. To strengthen the construct validity, we also predefined 6 hypotheses for convergent and discriminant validity. These hypotheses were formulated before the analyses were performed, and tested against existing measurements with known validity, which is a strength for construct validation (de Vet et al., 2011).

5.8 Reliability and measurement error

Test-retest reliability should be assessed in a stable population with an appropriate time interval between measurements (Terwee et al., 2007). The time interval between the measurements was median 3 days (range 1-10), which means shorter than recommended for many of the participants. There is a potential risk of recall bias if the interval between the test and the retest is too short. We believe though, that the comprehensive questionnaire with a high number of questions filled out at baseline most likely reduce recall bias when filled out only few days later. In addition, sensitivity analyses with exclusion of the 5 participants who reported change in the physical workload from first test to retest was conducted. The results revealed similar results in all reliability analysis, indicating the results to be reliable.

The PWQ scores are continuous measures, therefore ICC is the most suitable reliability parameter for relative reliability. However, there are several models to choose from, all suited for different situations (de Vet et al., 2011). As systematic differences are considered to be part of the measurement error (Terwee et al., 2010), the $ICC_{2.1}$ (two-way ANOVA random effect model for absolute agreement) was chosen for this study. This model makes it possible to generalize our reliability results (de Vet et al., 2011). ICC is a commonly used method to evaluate reliability, but it is insufficient to fully assess the reliability of measurements, therefore SEM and SDC were used to assessment of the measurement error.

To account for the systematic differences that can occur in self-reported questionnaires, the $SEM_{agreement}$ was used. Many researchers are calculating the SEM from the formula SEM = $SD\sqrt{1 - ICC}$. However, de Vet et al. (2011) are warning against this formula due to several fallacies. By deriving the SEM value from the error variance in the ICC formula, there is a risk of not including the systematic error. In the present study we wanted to make sure to include the systematic error, therefore SEM was estimated from the VARCOMP analysis. In this formula, the difference of the values of the two measurements are found, and the mean and the SD of these differences are calculated: $SEM_{agreement}(o^2{}_o+o^2{}_{po,\ e})$, where $o^2{}_o$ is the variance due to systematic error between observations and $o^2{}_{po,\ e}$ is the random error (de Vet et al., 2011). The SEM is presented in the same scale as the subscale scores, and therefore gives an easy picture of the size of the measurement error. To assess whether the change was a “real” individual change above the measurement error, also the $SDC_{95\%}$ind was calculated.

6 DISCUSSION OF RESULTS

6.1 Exploratory factor analysis

As a result of the EFA, a two-factor structure of the PWQ found to explain 58,3% of the total variance. A total of twenty -two items remained in the final questionnaire. The items “prolonged sitting”, “VDU work for long periods of time” and “climbing stairs” were deleted due to negative or too low loading on the factors. The item “handling peddles with feet” showed low communalities value, indicating a poor fit with the other items and were therefore deleted. The present result were in line with the original study of Bot et al. (2004) in terms of the number of factors obtained and also in terms of the nature of the items comprising each of the factors, which could be labelled; subscale 1 “Heavy physical workload” and subscale 2 “Long lasting postures and repetitive movements”. The results differed though in terms of number of items included in each subscale. In addition to the 12 items Bot et al. (2004) included in the subscale “heavy physical workload”, we found 3 more items to be included; “work with vibrating tools” and “walking on irregular surfaces” loaded strongly on this factor, and “twisted posture” loaded above 0.5 on both factors, but slightly stronger on factor 1 and were therefore obtained in subscale 1 (heavy physical workload). In subscale 2, Bot et al. (2004) included 6 items, whereas the present study also revealed that the item “uncomfortable posture” loaded above 0.5 and could be obtained in this subscale. The difference in number of items in each subscale was not entirely unexpected in light of the difference of the participant characteristics in the two studies. Furthermore, in the present study, back pain was the most frequently reported pain area. Back pain is in several previous studies found to be associated with risk factors such as twisted posture (Videman, Ojajarvi, Riihimaki, & Troup, 2005) (Van Nieuwenhuyse et al., 2006) (Campo, Weiser, Koenig, & Nordin, 2008), working with vibrating tools (Bovenzi, Schust, & Mauro, 2017) and uncomfortable postures/ awkward postures (Bovenzi, 2009) (Videman et al., 2005) (Van Nieuwenhuyse et al., 2006). This may explain why these items loaded strong enough to be included in the subscales in the present study. When it comes to “walking on irregular surfaces”, previous studies have indicated that uneven surfaces are more challenging for postural control which can cause problems when handling loads and requires the worker to compensate by placing the body in awkward positions or to put more force into the task (Gates, Scott, Wilken, & Dingwell, 2013; Marigold & Patla, 2008; Thies, Richardson, & Ashton-Miller, 2005). Hence, this item may also be associated with participants with low back pain.

6.2 Internal consistency

The subscale values of the PWQ had good internal consistencies, 0.94 for the "heavy physical workload" and 0.85 for the "long lasting postures and repetitive movements". This result is consistent with the original study of Bot et al. (2004), showing Cronbach's alpha of 0.92-0.93 on the "heavy physical workload" subscale and 0.86-0.87 on the "long lasting postures and repetitive movements" subscale (from three different study populations with upper or lower extremity complaint). The high values of internal consistency of both subscales are supporting the results from the factor analyses, indicating that the PWQ is an adequate reflection of the dimensionality of the construct physical workload, showing good structural validity.

6.3 Construct validity

For six of the eight first hypotheses there were enough evidence to confirm the hypotheses that the median values were statistically significant different ($p<0.01$) among occupational groups. Although the occupational group with heavy physical load scored higher on the subscale with "heavy physical workload" than the occupational group with no physical load, the difference was not statistically significant. Additionally, the occupational group with heavy physical load scored higher on the subscale with "heavy physical workload" than on the subscale with "long lasting postures and repetitive movements", but the difference was not statistically significant. As the factor analysis in the present study revealed different items than Bot et al. (2004) to be included in the two subscales, we must be careful with comparing our results with Bot et al. However, there might be a few similarities to point out anyway. In line with Bot et al. (2004) we also found that the questionnaire could clearly distinguish between the scores of the occupational group with long lasting postures and repetitive movements. This group scored low on the first subscale and high on the second, and all the hypotheses regarding occupations classified as "both heavy physical load and long-lasting postures and repetitive movements" were confirmed and further provides initial evidence for the validity of the questionnaire. The hypotheses regarding convergent validity and discriminant validity were confirmed. The MEI, which is a self-report questionnaire assessing mechanical exposure of the shoulder- neck region (Balogh et al., 2001), shows high correlation with the "heavy physical workload" subscale and moderate correlation with the "long lasting postures and repetitive movements" subscale. The SF-36 dimensions; "physical functions" and "general health" (Ware & Sherbourne, 1992), measure other constructs than physical workload, and as expected, low correlation was found with both subscales. The validity analyses confirmed 85% of the predefined hypotheses, indicating a good construct validity (Terwee et al., 2007).

6.4 Reliability and measurement error

The $ICC_{2.1}$ was well above the minimum standard of both subscales, 0.96 and 0.92 on subscale 1 and 2, respectively, and therefore considered to be good (de Vet et al., 2011). There was a decrease in difference score from test to re-test in both subscales: -5.1 ± 8.5 on subscale 1 and -6.5 ± 12.6 on subscale 2. Although this was a statistically significant result, it may be considered to be low as the scale ranged from 0-100. However, the absolute reliability, presented as measurement error and reported in the actual scale unit, is more clinically useful than the relative reliability. Although the ICC was high, the SEM of 6.9 and 10.0 and consequently $SDC_{95\%ind}$ of 19.2 and 27.7 on subscale 1 and 2, respectively, indicates that only large changes of workload may be identified by the PWQ. The $SDC_{95\%ind}$ indicates the smallest within-person change that can be interpreted as a “real” individual change above the measurement error, which indicates that a score of self- reported physical workload at the individual level would have to change by 19.2 and 27.7 on subscale 1 and 2, respectively, to ensure that the change was not a result of measurement error (de Vet et al., 2011). Since we were unable to identify other studies assessing the measurement properties of the PWQ, and Bot et al. (2004) only evaluated internal consistency and validity, we believe this is the first study to determine the test-retest reliability of the questionnaire, and we have no other studies to compare with. To determine whether the SEM and SDC values are acceptable depends on which changes are minimally important on the PWQ subscales. The SDC should be smaller than the minimal important change (MIC) (Terwee et al., 2007), however, no MIC for the PWQ scales are available, and we were unable to determine whether the SDC of the PWQ scales were sufficiently low. However, these relatively high SEM and SDC values are similar to quality of life outcomes on the SF-36 scale among patients with chronic neck pain (Juul, Søgaard, Davis, & Roos, 2016), and as the SF-36 has widely been used for its good measurement properties, this may indicate that our values are common on a 0-100 scale. Future studies are needed to address the responsiveness, i.e. the ability to measure change of the PWQ subscales.

6.5 Clinical and research implications

The present study reveals a questionnaire with acceptable measurement properties which can be used to assess physical workload in patients with various musculoskeletal disorders. The self- administrative questionnaire is short, simple, inexpensive and feasible to use in assessment of physical workload in several settings. We can recommend using the Norwegian version of the PWQ in assessment of physical workload exposure in clinical settings and occupational health, as screening in large epidemiological studies, or in smaller research purposes. In addition, the present study's results also make it possible to detect changes of exposure of physical workload in a group of individuals over time or after an intervention.

7 CONCLUSION

The Norwegian version of the PWQ, consisting of two subscales: "Heavy physical workload" (15 items) and "Long lasting postures and repetitive movements" (7 items), shows acceptable internal consistency, construct validity and reliability when used on patients with various musculoskeletal disorders receiving rehabilitation in a secondary outpatient clinic in Norway. This study indicates that the PWQ can be used in clinical and occupational healthcare and in research purposes on patients with musculoskeletal disorders. However, further research is advised on test-retest reliability and construct validity in other populations and clinical settings. Also, the clinical value of the PWQ in relation to work related musculoskeletal disorders should be investigated further.

REFERENCES

Aittomäki, A., Lahelma, E., Roos, E., Leino-Arjas, P., & Martikainen, P. (2005). Gender Differences in the Association of Age with Physical Workload and Functioning. *Occupational & Environmental Medicine, 62*(2), 95-100 DOI: https://doi.org/10.1136/oem.2004.014035

Althubaiti, A. (2016). Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare, 9*, 211-217 DOI: https://doi.org/10.2147/JMDH.S104807.

Andersen, L. L., Fallentin, N., Thorsen, S. V., & Holtermann, A. (2016). Physical workload and risk of long-term sickness absence in the general working population and among blue-collar workers: prospective cohort study with register follow-up. *Occupational & Environmental Medicine, 73*(4), 246-253 DOI: https://doi.org/10.1136/oemed-2015-103314.

Andresen, E. M. (2000). Criteria for assessing the tools of disability outcomes research. *Archives of Physical Medicine and Rehabilitation, 81*, 15-20 DOI: https://doi.org/10.1053/apmr.2000.20619.

Balogh, I., Ørbaek, P., Winkel, J., Nordander, C., Ohlsson, K., & Ektor-Andersen, J. (2001). Questionnaire-based mechanical exposure indices for large population-reliability, internal consistency and predictive validity. *Scandinavian Journal of Work, Environment and Health, 27*(1), 41-48 DOI: https://doi.org/10.5271/sjweh.585.

Balogh, I., Ørbæk, P., Ohlsson, K., Nordander, C., Unge, J., Winkel, J., & Hansson, G. Å. (2004). Self-assessed and directly measured occupational physical activities—influence of musculoskeletal complaints, age and gender. *Applied Ergonomics, 35*(1), 49-56 DOI: https://doi.org/10.1016/j.apergo.2003.06.001.

Bang, C. K., Lund, T., Labriola, M., Villadsen, E., & Bültmann, U. (2007). The fraction of long-term sickness absence attributable to work environmental factors: prospective results from the Danish Work Environment Cohort Study. *Occupational & Environmental Medicine, 64*, 487-489 DOI: https://doi.org/10.1136/oem.2006.028563.

Barrero, L. H., Katz, J. N., & Dennerlein, J. T. (2009). Validity of self-reported mechanical demands for occupational epidemiologic research of musculoskeletal disorders. *Scandinavian Journal Of Work, Environment & Health, 35*(4), 245-260 DOI: https://doi.org/10.5271/sjweh.1335.

Beaton, E. D., Bombardier, E. C., Guillemin, E. F., & Ferraz, E. M. B. (2000). Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures. *Spine, 25*(24), 3186-3191 DOI: https://doi.org/10.1097/00007632-200012150-00014.

Bot, S. D. M., Terwee, C., van Der Windt, D. A. W. M., Feleus, A., Bierma-Zeinstra, S. M., Knol, D. L., . . . Dekker, J. (2004). Internal consistency and validity of a new physical workload questionnaire. *Occupational & Environmental Medicine, 61*, 980-986 DOI: https://doi.org/10.1136/oem.2003.011213.

Bovenzi, M. (2009). Metrics of whole-body vibration and exposure-response relationship for low back pain in professional drivers: a prospective cohort study. *Int Arch Occup Environ Health, 82*(7), 893-917 DOI: https://doi.org/10.1007/s00420-008-0376-3.

Bovenzi, M., Schust, M., & Mauro, M. (2017). An overview of low back pain and occupational exposures to whole-body vibration and mechanical shocks. *La Medicina del lavoro, 108*, 419-433 DOI: https://doi.org/10.23749/mdl.v108i6.6639.

Burdorf, A., & Van Der Beek, A. J. (1999). In musculoskeletal epidemiology are we asking the unanswerable in questionnaires on physical load? *Scandinavian Journal Of Work, Environment & Health, 25*(2), 81-83 DOI: https://doi.org/10.5271/sjweh.409.

Campo, M., Weiser, S., Koenig, K. L., & Nordin, M. (2008). Work-related musculoskeletal disorders in physical therapists: a prospective cohort study with 1-year follow-up. *Phys Ther, 88*(5), 608-619 DOI: https://doi.org/10.2522/ptj.20070127.

Carter, R. E., & Lubinsky, J. (2015). *Rehabilitation research: principles and applications* (5 ed.). St. Louis, Missouri: Elsevier.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research, 1*(2), 245-276 DOI: https://doi.org/10.1207/s15327906mbr0102_10.

Da Costa, B. R., & Vieira, E. R. (2010). Risk factors for work-related musculoskeletal disorders: a systematic review of recent longitudinal studies. *American journal of industrial medicine, 53*, 285-323 DOI: https://doi.org/doi:10.1002/ajim.20750.

Dale, A. M., Strickland, J., Gardner, B., Symanzik, J., & Evanoff, B. (2010). Assessing Agreement of Self-reported and Observed Physical Exposures of the Upper Extremity. *International Journal of Occupational Environmental Health, 16*(1), 1-10 DOI: https://doi.org/10.1179/107735210800546227.

David, G. C. (2005). Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders. *Occupational Medicine, 55*(3), 190-199 DOI: https://doi.org/10.1093/occmed/kqi082.

de Vet, H., Terwee, C., Mokkink, L., & Knol, D. (2011). *Measurement in Medicine: A Practical Guide*. doi:https://doi.org/10.1017/CBO9780511996214

Demarchi, S., Oliveira, C., Franco, M., Morelhão, P., Hisamatsu, T., Silva, F., . . . Pinto, R. (2019). Association of perceived physical overload at work with pain and disability in patients with chronic non-specific low back pain: a 6-month longitudinal study. *European Spine Journal, 28*(7), 1586-1593 DOI: https://doi.org/10.1007/s00586-019-05986-3.

Gates, D. H., Scott, S. J., Wilken, J. M., & Dingwell, J. B. (2013). Frontal plane dynamic margins of stability in individuals with and without transtibial amputation walking on a loose rock surface. *Gait & Posture, 38*(4), 570-575 DOI: https://doi.org/10.1016/j.gaitpost.2013.01.024.

Griffith, L. E., Shannon, H. S., Wells, R. P., Walter, S. D., Cole, D. C., Côté, P., . . . Langlois, L. E. (2012). Individual participant data meta-analysis of mechanical workplace risk factors and low back pain. *American journal of public health, 102*(2), 309-318 DOI: https://doi.org/10.2105/AJPH.2011.300343.

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology, 46*(12), 1417-1432 DOI: https://doi.org/10.1016/0895-4356(93)90142-N.

Hansson, G.-Å., Balogh, I., Byström, J., Ohlsson, K., Nordander, C., Asterland, P., . . . Skerfving, S. (2001). Questionnaire versus direct technical measurements in assessing postures and movements of the head, upper back, arms and hands. *Scandinavian Journal Of Work, Environment & Health*(1), 30-40 DOI: https://doi.org/10.5271/sjweh.584.

Hildebrandt, V. H., Bongers, P. M., van Dijk, F. J. H., Kemper, H. C. G., & Dul, J. (2001). Dutch Musculoskeletal Questionnaire: description and basic qualities. *Ergonomics, 44*(12), 1038-1055 DOI: https://doi.org/10.1080/00140130110087437.

Hollmann, S., Klimmer, F., Schmidt, K.-H., & Kylian, H. (1999). Validation of a Questionnaire for Assessing Physical Work Load. *Scandinavian Journal Of Work, Environment & Health, 25*(2), 105-114 DOI: https://doi.org/10.5271/sjweh.412.

Hubbard, R., & Allen, S. J. (1987). An empirical comparison of alternative methods for principal component extraction. *Journal of Business Research, 15*(2), 173-190 DOI: https://doi.org/10.1016/0148-2963(84)90047-X.

Ihlebæk, C., Brage, S., Natvig, B., & Bruusgaard, D. (2010). Forekomst av muskel- og skjelettlidelser i Norge. *Tidsskrift for Den norske legeforening, 130*(23), 2365-2368 DOI: https://doi.org/10.4045/tidsskr.09.0802.

Juul, T., Søgaard, K., Davis, A. M., & Roos, E. M. (2016). Psychometric properties of the Neck OutcOme Score, Neck Disability Index, and Short Form–36 were evaluated in patients with neck pain. *Journal of Clinical Epidemiology, 79*, 31-40 DOI: https://doi.org/10.1016/j.jclinepi.2016.03.015.

Kinge, J. M., Knudsen, A. K., Skirbekk, V., & Vollset, S. E. (2015). Musculoskeletal disorders in Norway: prevalence of chronicity and use of primary and specialist health care services. *BMC Musculoskeletal Disorders, 16*(1) DOI: https://doi.org/10.1186/s12891-015-0536-z.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases, 38*(1), 27-36 DOI: https://doi.org/10.1016/0021-9681(85)90005-0.

Kline, P. (1993). *The Handbook of psychological testing*. London: Routledge.

Leijon, O., Wiktorin, C., Härenstam, A., & Karlqvist, L. (2002). Validity of a self administered questionnaire for assessing physical work loads. *Journal Of Occupational And Environmental Medicine, 44*(8), 724-735 DOI: https://doi.org/10.1097/01.jom.0000026044.24145.9e.

Marigold, D. S., & Patla, A. E. (2008). Age-related changes in gait for multi-surface terrain. *Gait & Posture, 27*(4), 689-696 DOI: https://doi.org/10.1016/j.gaitpost.2007.09.005.

Mayer, J., Kraus, T., & Ochsmann, E. (2012). Longitudinal evidence for the association between work-related physical exposures and neck and/or shoulder complaints: a systematic review. *International Archives of Occupational and Environmental Health, 85*(6), 587-603 DOI: https://doi.org/10.1007/s00420-011-0701-0.

Mehlum, I. S., Kristensen, P., Kjuus, H., & Wergeland, E. (2008). Are Occupational Factors Important Determinants of Socioeconomic Inequalities in Musculoskeletal Pain? *Scandinavian Journal Of Work, Environment & Health, 34*(4), 250-259 DOI: https://doi.org/10.5271/sjweh.1269.

Mokkink, L., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., . . . de Vet, H. (2010a). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research, 19*(4), 539-549 DOI: https://doi.org/10.1007/s11136-010-9606-8.

Mokkink, L., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., . . . de Vet, H. (2010b). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737-745 DOI: https://doi.org/10.1016/j.jclinepi.2010.02.006.

National Research Council Institute of Medicine. (2001). *Musculoskeletal Disorders and the Workplace: Low Back and Upper Extremities*. doi:https://doi.org/10.17226/10032

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*: McGraw-Hill.

Pallant, J. (2016). *SPSS survival manual: a step by step guide to data analysis using IBM SPSS* (6 ed.). Maidenhead: McGraw Hill.

Pope, D. P., Silman, A. J., Cherry, N. M., Pritchard, C., & Macfarlane, G. J. (1998). Validity of a self-completed Questionnaire Measuring the Physical Demands of Work. *Scandinavian Journal Of Work, Environment & Health, 24*(5), 376-385 DOI: https://doi.org/10.5271/sjweh.358.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3 ed.). Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Sterud, T. (2014). Work-related mechanical risk factors for long-term sick leave: a prospective study of the general working population in Norway. *The European Journal of Public Health, 24*(1), 111-116 DOI: https://doi.org/10.1093/eurpub/ckt072.

Sterud, T., & Tynes, T. (2013). Work-related psychosocial and mechanical risk factors for low back pain: a 3-year follow-up study of the general working population in Norway. *Occupational & Environmental Medicine, 70*, 296-302 DOI: https://doi.org/10.1136/oemed-2012-101116.

Stock, S. R., Fernandes, R., Delisle, A., & Vezina, N. (2005). Reproducibility and validity of workers' self-reports of physical work demands. *Scandinavian Journal Of Work, Environment & Health, 31*(6), 409-437 DOI: https://doi.org/10.5271/sjweh.947.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6 ed.). Harlow: Pearson.

Takala, E. P., Pehkonen, I., Forsman, M., Hansson, G.-Å., Mathiassen, S. E., Neumann, W. P., . . . Winkel, J. (2010). Systematic evaluation of observational methods assessing biomechanical exposures at work. *Scandinavian Journal Of Work, Environment & Health, 36*(1), 3-24 DOI: https://doi.org/10.5271/sjweh.2876.

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van Der Windt, D. A. W. M., Knol, D. L., Dekker, J., . . . de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34-42 DOI: https://doi.org/10.1016/j.jclinepi.2006.03.012.

Terwee, C. B., Mokkink, L., Poppel, M., Chinapaw, M., Mechelen, W., & Vet, H. (2010). Qualitative Attributes and Measurement Properties of Physical Activity Questionnaires. *Sports Medicine, 40*(7), 525-537 DOI: https://doi.org/10.2165/11531370-000000000-00000.

Thies, S. B., Richardson, J. K., & Ashton-Miller, J. A. (2005). Effects of surface irregularity and lighting on step variability during gait: A study in healthy young and older women. *Gait & Posture, 22*(1), 26-31 DOI: https://doi.org/10.1016/j.gaitpost.2004.06.004.

Tremblay, M. S., Colley, R. C., Saunders, T. J., Healy, G. N., & Owen, N. (2010). Physiological and health implications of a sedentary lifestyle. *Applied Physiology, Nutrition, and Metabolism, 35*(6), 725-740 DOI: https://doi.org/10.1139/H10-079.

Tynes, T., Aagestad, C., Andersen, L., Perkio-Makela, M., Pinilla Garcia, F., Vermeylen, G., . . . Formazin, M. (2017). Physical working conditions as covered in European monitoring questionnaires. *BMC Public Health, 17*, 1-9 DOI: https://doi.org/10.1186/s12889-017-4465-7.

van Der Beek, A. J., & Frings-Dresen, M. H. (1998). Assessment of mechanical exposure in ergonomic epidemiology. *Occupational & Environmental Medicine, 55*(5), 291-299 DOI: https://doi.org/10.1136/oem.55.5.291.

Van Nieuwenhuyse, A., Somville, P. R., Crombez, G., Burdorf, A., Verbeke, G., Johannik, K., . . . Moens, G. F. (2006). The role of physical workload and pain related fear in the development of low back pain in young workers: evidence from the BelCoBack Study; results after one year of follow up. *Occupational & Environmental Medicine, 63*(1), 45-52 DOI: https://doi.org/10.1136/oem.2004.015693.

Videman, T., Ojajarvi, A., Riihimaki, H., & Troup, J. D. (2005). Low back pain among nurses: a follow-up beginning at entry to the nursing school. *Spine, 30*(20), 2334-2341 DOI: https://doi.org/10.1097/01.brs.0000182107.14355.ca.

Viikarijuntura, E., Rauas, S., Martikainen, R., Kuosma, E., Riihimaki, H., Takala, E.-P., & Saarenmaa, K. (1996). Validity of self-reported physical work load in epidemiologic studies on musculoskeletal disorders. *Scandinavian Journal of Work Environmen & Health, 22*(4), 251-259 DOI: https://doi.org/10.5271/sjweh.139.

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30*(6), 473-483. Retrieved from http://www.jstor.org/stable/3765916

Watkins, M. (2000). Monte Carlo PCA for Parallel Analysis (Computer Software). *State College, PA: Ed & Psych Associates*.

Wiktorin, W. C., Hjelm, W. E., Winkel, J., & Köster, M. (1996). Reproducibility of a Questionnaire for Assessment of Physical Load During Work and Leisure Time. *Journal of Occupational & Environmental Medicine, 38*(2), 190-201 DOI: https://doi.org/10.1097/00043764-199602000-00017.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin, 99*(3), 432-442 DOI: https://doi.org/10.1037/0033-2909.99.3.432

ARTICLE DRAFT

MEASUREMENT PROPERTIES OF THE NORWEGIAN VERSION OF THE PHYSICAL WORKLOAD QUESTIONNAIRE (PWQ)

Author: Lise Grethe Kjønø

Supervisor: Kjersti Storheim

Co-supervisor: Ørjan Nesse Vigdal

Journal for submission: BMC Public Health

https://bmcpublichealth.biomedcentral.com/submission-guidelines/preparing-your-manuscript/research-article

Research article

MEASUREMENT PROPERTIES OF THE NORWEGIAN VERSION OF THE PHYSICAL WORKLOAD QUESTIONNAIRE (PWQ):

A CROSS-SECTIONAL STUDY WITH A TEST-RETEST DESIGN

Author: Lise Grethe Kjønø

More co-authors will be included before submitting the article.

Abstract

Background: The Physical Workload Questionnaire (PWQ) is a questionnaire based on an item pool of 26 items related to physical workload. The PWQ has recently been translated into Norwegian, but its measurement properties have not yet been tested. The aim of this study is to assess the measurement properties of the Norwegian version of the PWQ in patients with musculoskeletal disorders.

Method: A cross-sectional study with a test-retest design were conducted to assess factorial structure, internal consistency, construct validity and reliability in patients with musculoskeletal disorders. Exploratory Factor Analysis (EFA) was performed to assess the factorial structure and number of items to be included in the Norwegian version of the PWQ. Internal consistency was assessed using Cronbach's alpha, and construct validity by 14 a priori hypotheses ("known" group, convergent and discriminant validity). Reliability was evaluated by Intraclass Correlation Coefficient ($ICC_{2.1}$), Standard Error of Measurement ($SEM_{agreement}$) and Smallest Detectable Change ($SDC_{95\%}$ind).

Results: A total of 115 patients with musculoskeletal disorders and a mean age (SD) of 46 (9) years were included in the cross-sectional study, of which 48 were included in the reliability analyses. EFA on the item pool resulted in two subscales; "Heavy physical workload" (15 items, range 0-100) and "Long lasting postures and repetitive movements" (7 items, range 0-100). The internal consistency with Cronbach's alpha was 0.94 and 0.85 on subscale 1 and 2, respectively. There were no floor or ceiling effects of the subscales. Assessment of construct validity showed that 12 of the 14 (85%) predefined hypotheses were confirmed. Reliability of PWQ data demonstrated an $ICC_{2.1}$ of 0.96 (95% CI 0.88, 0.98) and 0.92 (95% CI 0.81, 0.96), SEM of 6.9 and 10.0 and $SDC_{95\%}$ind of 19.2 and 27.7 of subscale 1 and 2, respectively.

Conclusion: The Norwegian version of the PWQ demonstrated acceptable internal consistency, construct validity and reliability and can be used to assess physical workload in patients with various musculoskeletal disorders.

Keywords: Physical workload questionnaire, PWQ, physical workload, musculoskeletal disorders, validity, reliability

Background

Physical work demands are thought to be associated with development of various musculoskeletal disorders (1, 2), and to be related to labour market participation (3, 4). Work-related musculoskeletal disorders are associated with sickness absence in a variety of occupations, and are among the leading causes of sickness absence in Europe (5, 6). According to existing literature, the major physical risk factors for work-related musculoskeletal disorders seems to be heavy lifting, working with a bent or twisted back or elevated arms, repetitive movements and vibration (1, 2).

To prevent and reduce work-related musculoskeletal disorders, it is necessary to assess physical workload at the workplace (7). Investigating the contribution of physical workload to musculoskeletal disorders may be based on direct measurements, observational methods or self-reported questionnaires (8). Direct measurements and observational methods are very resource demanding, especially in large epidemiological studies, therefore self-reported questionnaires are commonly used and offer the possibility of studying a great number of persons at a modest cost, as well as allowing the investigation of a large number of variables when collecting exposure information (7). However, numerous studies have warned researchers about the lack of validity of self-reported questionnaires (8, 9).

A wide variety of questionnaires have been developed to assess physical workload, but often their measurement properties have not been properly tested (7, 10). The Physical Workload

Questionnaire (PWQ) was developed by Bot et al., with the aim to create a short and simple self-report questionnaire for assessing physical workload and to be used both in occupational health care as well as in epidemiological research (11). The items were captured from the Dutch Musculoskeletal Questionnaire (DMQ) which is a screening instrument for musculoskeletal workload and associated potentially hazardous working conditions (12). The items in the DMQ were based on reviews of the epidemiological literature. In contrast to other physical workload questionnaires, the DMQ has been studied thoroughly for its dimensionality and validity (12). However, the domain taken into account regarding musculoskeletal workload has 63 items and is therefore too lengthy to be used in most studies. The selection of the items for the PWQ were based on face validity and a discussion among experts during a consensus meeting. An item pool of 26 items that were expected to have an association with either upper or lower extremity complaints were chosen and tested for its dimensionality, internal consistency and construct validity in a population with upper extremity and lower extremity musculoskeletal disorders in the Netherlands. The item pool formed two subscales and the results supported the internal structure, internal consistency and construct validity (11).

To our knowledge the PWQ has not yet been translated to other languages or been tested for its measurement properties in other studies. Before a measurement instrument can be used in research or clinical practice, its measurement properties should be assessed and considered adequate (13). The PWQ item pool is previously translated and cross-culturally adapted into Norwegian according to international guidelines. To enable the use of the Norwegian version of the PWQ in occupational health care, clinical settings and research in Norwegian speaking patients, the present study aims to test its dimensionality and measurement properties in terms of internal consistency, construct validity and reliability in patients with musculoskeletal disorders.

Methods

This methodological study is part of the BACE (BAck Complaints in Elders) study in Norway. BACE is a prospective cohort study designed to assess elderly (≥55 years) patients with back pain. The PWQ is part of a comprehensive questionnaire used in the cohort study. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist was applied in the design of the study (13). The PWQ will be assessed according to guidelines for Patient-Related Outcome-Measures (PROMs) (14).

Design

This is a cross-sectional study including a test-retest design to test the Norwegian version of the PWQ for its measurement properties.

Participants

Participants were recruited from secondary care, at an outpatient rehabilitation clinic in Akershus, Norway, between November 2015 and January 2018. Eligible participants were patients with musculoskeletal disorders, aged 18 or above, working or on sick leave. Exclusion criteria were patients being unable to speak, read or write in Norwegian. Inclusion was performed by a clinician, primarily a physiotherapist, meeting the patient at the clinic. At baseline all patients received written and oral information about the study, and signed informed consent was obtained from all patients. The study is considered a quality assessment project by the Regional Ethical Committees for Medicine and Health (REC) (reference no. 2014/1634/REK vest) and was approved by the Norwegian Centre for Research Data (NSD) (reference No. 42149) in 2019.

According to recommended quality criteria by Terwee et al. (14) and Nunnally (15) it was planned to recruit 100 patients. These criteria suggest a minimum of 100 participants for assessing internal consistency, at least 50 participants for assessing reproducibility and floor or ceiling effects (14), and at least 4-10 participants for each item being included in factor analysis (15).

Procedures and measurements

At baseline (the cross-sectional study) the included participants completed the PWQ items as part of a comprehensive questionnaire, which also included sociodemographic variables, pain localization, intensity and history, psychosocial work environment, health-related productivity costs and health-related quality of life. A drawing map were used to measure pain localisation (16). The patients marked the area or areas that had been painful during the last week, and on the basis of their marking they were categorized in the following body areas of pain: lower extremities, back, neck, shoulder/upper extremities and head. The Numeric rating scale (NRS) were used to measure pain intensity (17). The NRS is a subjective measure in which the participants were asked to rate their pain during the last week on an 11 point-numerical scale. The scale is composed of 0 (no pain at all) to 10 (worst possible pain). The General Nordic Questionnaire for psychological and social factors at work (QPSnordic) were used to measure characteristics of the psychosocial work environment (18) and questions from and the iProductivity Cost Questionnaire (iPCQ) to measure health-related productivity costs (19). The iPCQ is measuring productivity losses of paid work due to absenteeism, presenteeism and productivity losses related to unpaid work (19). In the present study we used questions from the iPCQ in assessment of work status; occupation, paid job, working days/hours a week, sick leave and rehabilitation/work disability. Health-related quality of life was measured by The short form 36 Health status questionnaire (SF-36). The SF-36 comprises eight different health dimensions, in the present study we used two: physical function (10 items) and general health (5 items). The patients' responses were converted to a scale ranging from 0 (worst possible score) to 100 (best possible score) (20). In addition, a questionnaire, The Mechanical Exposure Index (MEI) was used to assess physical workload (21). MEI is a self-report questionnaire assessing mechanical exposure of the shoulder- neck region. The questionnaire consists of twelve questions based on awkward body postures, static load and work involving repetitive and precision movements, manual material handling and vibration. On the basis of 3 response alternatives; 0 (nothing/hardly nothing), 1 (somewhat), 2 (a great deal), an index was calculated ranging from 0-24 (21).

Patients consenting to participate at the reliability part of the study filled out the PWQ at their second attendance, preferably 1 week after they first filled out the questionnaire. Additionally, a global question recording change in the work condition in the time interval was completed. Patients reporting “unchanged” in working conditions were considered stable and included in test-retest reliability analysis.

The Physical Workload Questionnaire

The PWQ is a self-report questionnaire for assessing physical workload in occupational healthcare and in research (11). The questionnaire was based on an item pool consisting of 26 items assessing force, dynamic and static load, repetitive load, (uncomfortable) postures, sitting, standing and walking. In the only former study, the item pool was assessed for its dimensionality, internal consistency and construct validity in a population with upper extremity and lower extremity musculoskeletal disorders. Factor analysis revealed two subscales; twelve items related to the first subscale “heavy physical work” and six items related to the second subscale “long lasting postures and repetitive movements”. The remaining eight items (“prolonged sitting”, “prolonged visual display units work”, “work with vibrating tools”, “operate peddles with feet”, “climbing stairs”, “twisted posture”, “uncomfortable postures” and “walking on irregular surfaces”) were excluded due to low loadings or to similar loadings on both factors. Both subscales showed good internal consistency evaluated by Cronbach’s alpha (11). The construct validity was examined based on hypotheses that physical workload would vary among different occupational groups, and

Bot et al. classified the occupations of the participants in their study into four groups on the basis of expected workload; “heavy physical load”, “long lasting postures and repetitive movements”, both heavy physical load and long lasting postures and repetitive movements” and “no physical load”. Six of eight hypotheses regarding the construct validity of the subscales were confirmed, and the validity of the questionnaire developed by Bot et. al. was considered to be good. Each item is scored on a 4-point Likert scale with the response options; “seldom or never” (0), “sometimes” (1), “often” (2), and “(almost) always” (3). Scoring is done by adding up the response to each item, which produces a raw score. The final scores are calculated by dividing the raw score by the maximum score possible on the subscale, multiplied by 100, resulting in a final score ranging between 0 (no workload) and 100 (highest workload) for each subscale (11). The Norwegian version of the 26 items of the PWQ is shown in additional file 1.

Analysis

All data analyses were performed using SPSS version 26 (SPSS Inc., Chicago, IL). Demographic data were described using descriptive statistics. Numerical variables were expressed as means and standard deviation (SD) for normally distributed variables while median and minimum-maximum values were used for non-normally distributed variables. Categorical variables were expressed as frequencies. If one or more variables were skewed, non-parametric analysis were performed (22).

Exploratory factor analysis

The factorial structure of the PWQ was explored using exploratory factor analysis (EFA) based on the same 26 item pool forming the basis of the study of Bot et al. The suitability of data for factor analysis was confirmed using the Kaiser-Meyer-Olkin measure of sampling adequacy (values above 0.6 considered acceptable), a significant Bartlett`s Test of Sphericity and inspection of the correlation matrix (correlation coefficients of .3 and above preferable) (23).

Principal component analysis (PCA) was used to extract the factors followed by oblique rotation of factors using oblimin rotation. The number of factors to be retained was guided by three decision rules: Kaiser´s criterion, retention of eigenvalues above 1, Cattel´s scree plot (24), and by the use of Horn´s parallel analysis. Parallel analysis was conducted using the Monte Carlo PCA software developed by Watkins (25) that involves comparison of eigenvalues from the PCA with a set derived from a randomly generated datafile with the same number of items and cases. Factors with eigenvalues exceeding those obtained from the random datafile are suggested retained (25).

To aid in the interpretation of the retained factors, we computed factor loadings after direct oblimin rotation, allowing factors to correlate (23). Next step involved interpreting the rotated solution by identifying which items loaded on each retained factor. Items with factor loading below 0.5 (15), and communalities value below 0.3 were excluded (22). Items which cross-loaded were retained in the factor it loaded most strongly. A subscale in a questionnaire should be comprised of least three items (26).

Internal consistency

The internal consistency of the subscales was examined using Cronbach's alpha. Cronbach's alpha between 0.70 and 0.95 gave a positive rating (14). The item-total correlation was examined and items with values below 0.3 was considered excluded (26).

Data quality

Proportions of missing data and floor and/or ceiling effects were described. Floor or ceiling effects were considered to be present if more than 15% of the patients reported the lowest or the highest possible score, respectively (14).

Construct validity

The COSMIN group recommend to create predefined hypotheses to calculate the construct validity (14). Construct validity was assessed by “known” group validity, convergent validity and discriminant validity. “Known” group validity was tested with the same procedure as in the original study of Bot et al., where it was hypothesised that physical workload would vary among different occupational groups. The occupations of all included subjects were classified into four groups on the basis of expected physical load and the subscale scores of the occupational groups were compared.

- Group 1: no physical load (for example teacher, manager)
- Group 2: heavy physical load (for example nurse, child care worker)
- Group 3: long lasting postures and repetitive movements (for example cashier, civil servant, engineer)
- Group 4: both heavy physical load and long lasting postures and repetitive movements (for example electrician, agriculturer, mechanic)

Two investigators made the classification independently. Disagreements were discussed during a consensus meeting. Occupations which could not be agreed on were further discussed and resolved with a third investigator. Three occupations could not be classified (police, shop assistant and service employee) and participants with such occupations were therefore excluded from the construct validity analyses.

In addition, we predefined six hypotheses based on other health-related constructs. To assess convergent validity, both subscales were validated against the MEI (21), which covers many similar questions as in the PWQ, especially in the heavy physical workload subscale. We therefore expected high correlation between the “heavy physical workload” subscale and moderate to high correlation between the “long lasting postures and repetitive movements” subscale. To assess discriminant validity of the PWQ subscales, we formulated hypotheses about two of the dimensions in SF-36; “physical function” and “general health” (20). These dimensions measure another construct than the PWQ, we therefore expected low correlation between both of the subscales and the SF-36 dimensions. If more than 75% of the predefined hypotheses are confirmed, it can be concluded that the construct validity is good (14). The first eight hypotheses were rejected at significance levels of 1% ($p < 0.01$). Correlation coefficients under 0.3, between 0.3 and 0.6 and over 0.6 were considered low, moderate and high, respectively (27). All predefined hypotheses are listed in table 6.

Reliability and measurement error

A paired t-test was used to assess the mean difference between test and retest. An intraclass correlation coefficient ($ICC_{2,1}$) was used to assess relative reliability, and standard error of measurement (SEM) and smallest detectable change (SDC) were used to analyse absolute reliability (measurement error). $ICC_{2.1}$ and $SEM_{agreement}$ were used to account for the systematic difference between test and retest (26). SEM was estimated from the SPSS VARCOMP analysis; $SEM_{agreement} = \sqrt{(o^2_o + o^2_{po,\,e})}$, where o^2_o is the variance due to systematic error between observations and $o^2_{po,\,e}$ is the random error. Based on this, the SDC was estimated using the formula $SDC_{95\%}ind = 1.96 \times \sqrt{2} \times SEM_{agreement}$ (26). ICC of ≥ 0.70 was considered acceptable (14).

Results

A total of 115 patients with a mean (SD) age of 46 (9) were included in the cross-sectional study. The majority of the participants were women (68.7%), most of the included participants (90%) were in paid work and the working hours per week were median 37.5 hours (range 7.5-52 hours). On average they reported moderate pain, the majority had pain for more than 3 months, the most frequently reported pain area was the back region and physical workload were in general low. Sixty-two participated in the test- retest study of which 48 reported no change in working conditions from test to retest and had complete PWQ scores and could be included in the analysis. Patients participating in the test (n=115) and the re-test (n=48) were similar, but individuals included in the re-test had slightly different pain site locations and slightly different physical function and general health on SF-36. The time interval between test and re-test was median 3 days (range 1-10 days). Study sample characteristics are presented in Table 1.

Exploratory factor analysis

Inspection of the correlation matrix revealed the presence of many coefficients of 0.3 and above, Bartlett`s Test of Sphericity was highly significant (p=0.000) and the Kaiser-Meyer-Olkin measure of sampling adequacy value of 0.86, supported the factorability of the correlation matrix (23).

PCA revealed the presence of five factors with eigenvalues exceeding 1 (Kaiser`s criterion), explaining 38,9%, 16,3%, 5,9%, 4,1% and 4,1% of the variance respectively. However, the results of Horn's parallel analysis showed only two factors with eigenvalues exceeding the corresponding criterion values for a randomly generated data matrix of the same size (26 variables x 115 respondents), indicating that only two factors were appropriate for retention. The scree plot suggested either a two or three factor solution (figure 1), therefore both the three- factor and the two-factor solutions were inspected.

Table 1 Patient demographic characteristics and clinical status

	Cross sectional study sample (*n = 115)*		Test-retest study sample *(n=48)*	
	n=		**n=**	
Female gender (%)	115	68.7	48	62.5
Age in years, mean (SD)	115	45.6 ± 9.3	48	46.2 ± 8.6
Mother tongue Norwegian (%)	115	87	48	89.6
Educational level,	115		48	
Basic school level (%)		2.6		2.1
Upper secondary education (%)		39.1		35.4
University (%)		36.5		39.6
Other (%)		21.7		22.9
Work status				
Employed or self-employed (paid job), yes (%)	115	90.4	48	87.5
Working hours per week, median (range)	111	37.5 (7.5-52)	47	37.5 (7.5-52)
Working days per week, median (range)	102	5 (2-7)	41	5 (2-6)
Sick leave during last 4 weeks, yes (%)	114	68.7	48	68.8
Rehabilitation, work disability, yes (%)	115	6.1	48	8.4
Pain site, yes (%)				
Lower extremities		60.9		56.3
Back		69.6		75.0
Neck		49.6		35.4
Shoulder/upper extremities		47.8		31.3
Head		16.5		16.7
Pain duration in months, median (range)	89	0.16- 360	38	0.23- 360
< 3 months (%)		14.6		13.2
> 3 months (%)		85.4		86.8
Pain severity last week (NRS 0-10), mean (SD)	112	5.2 ± 2.0	48	4.9 ± 2.2
Physical Workload Questionnaire (PWQ 0-100)				
Heavy physical workload, median (range)	107	15.6 (0-82.2)	48	16.7 (2.2-82.2)
Long lasting postures and repetitive movements, mean (SD)	110	49.1 ±25.6	48	49.2 ±25.4
Mechanical exposure index (MEI 0-24), median (range)	111	4 (0-20)	47	4 (0-20)
Health-related quality of life (SF-36 0-100), median (range)				
Physical function	115	75 (30-100)	48	70 (30-95)
General health	115	57 (5-97)	48	68.5 (15-97)

NRS: Numeric rating scale (0 = no pain, 10 = worst possible pain), PWQ: Physical Workload Questionnaire (0 = no workload, 100 = highest workload), MEI: Mechanical Exposure Index (0 = no workload, 24 = highest workload), SF-36: The short form 36 Health status questionnaire (0 = maximum disability, 100 = no disability).

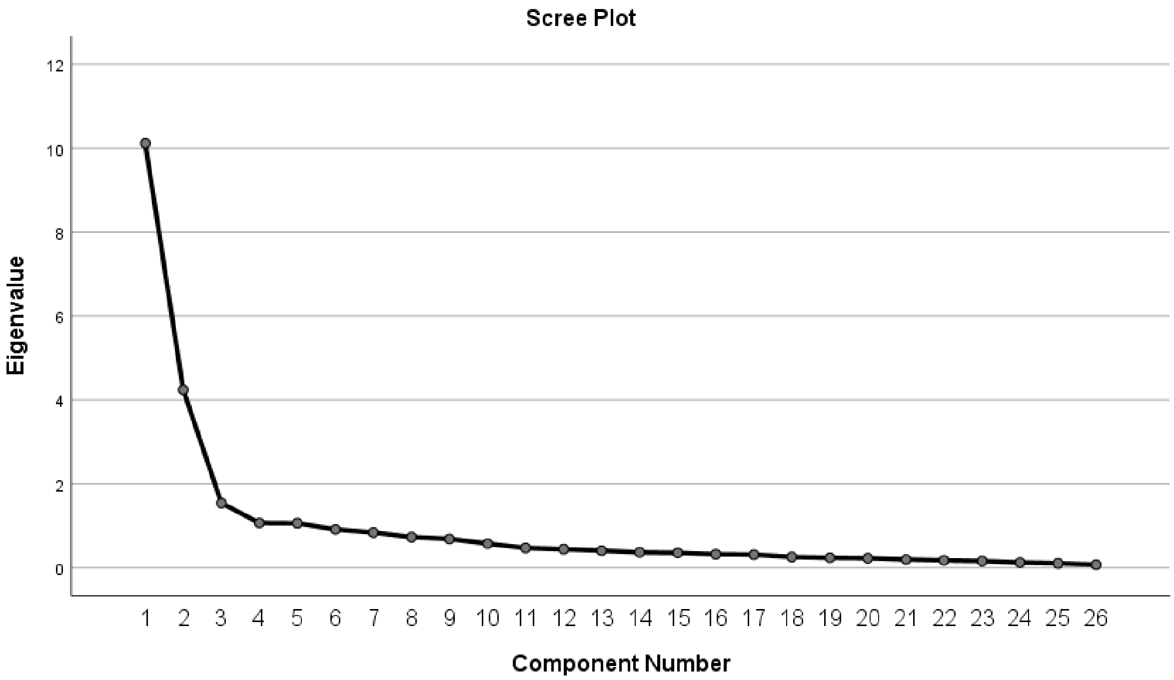

Figure 1 Scree plot of eigenvalues from the 26-item questionnaire applied in the population with various musculoskeletal disorders.

The three-factor solution explained a total of 61,2% of the variance, with factor 1 contributing 38,9%, factor 2 contributing 16,3% and factor 3 contributing 5,9% (table 2). Examination of the factor loadings revealed 9-15 items in factor 1 (depending on interpretation of pattern or structure matrix), 8 items in factor 2, but only two items to be retained in factor 3. The items "sitting" and "prolonged visual display units (VDU) work" loaded strongly on factor 3. These items were strongly intercorrelated, but showed negatively or very low correlations with other PWQ items, except from one item (item 18). Item 18 ("static posture") loaded just above 0.5 on factor 3, but it loaded stronger on the second factor and were therefore retained in factor two. A subscale in a questionnaire should be comprised of least three items (26), hence the three-factor solution was rejected.

Table 2 Pattern and Structure matrix of three-factor solution after EFA with oblimin rotation

Item	Pattern coefficients			Structure coefficients		
Does your work involve	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 1**	**Factor 2**	**Factor 3**
24. Walking on irregular surfaces	**0.84**			**0.78**		
25. Sitting/moving on knees	**0.76**			**0.79**		-0.39
22. Climbing stairs	**0.74**			**0.59**		
12. Hands below knees	**0.66**			**0.77**		-0.48
21. Handling peddles with feet	**0.62**			**0.61**		
14. Moving loads (>25kg)	**0.62**			**0.74**	0.34	-0.47
23. Often squatting	**0.61**			**0.72**		-0.55
13. Moving loads (>5kg)	**0.57**			**0.75**	0.39	-0.57
17. Physical hard work	**0.55**	0.35		**0.76**	**0.50**	-0.57
11. Hands above shoulders	0.45	0.32	-0.39	**0.67**	0.46	-0.57
5. Kneeling/squatting	0.42		-0.36	**0.62**	0.34	-0.56
20. Work with vibrating tools	0.38			**0.53**		-0.44
10. Wrists bent or twisted		**0.77**			**0.76**	
6. Repetitive movement		**0.77**	0.39		**0.74**	
8. Neck bent forward		**0.74**			**0.72**	
9. Turning/bending neck		**0.73**			**0.74**	
18. Static posture		**0.66**	**0.52**		**0.61**	**0.53**
7. Twisted posture		**0.65**		0.46	**0.72**	-0.44
19. Uncomfortable posture		**0.62**		0.37	**0.70**	
26. Repetitive tasks arms/hands		0.49	0.45		**0.57**	0.39
15. Exert force with arms	0.42	0.47		**0.66**	**0.61**	-0.53
16. Maximal force exertions	0.39			**0.60**	**0.58**	-0.48
2. Sitting			**0.87**	-0.40		**0.88**
3. VDU work			**0.83**	-0.43		**0.86**
1. Standing			-0.80	0.38		-0.81
4. Walking	0.40		-4.47	**0.61**		-0.64
Variance explained†	38.9%	16.3%	5.9%			
Total variance explained†	61.2%					

Factor loadings <0.3 are removed
Factor loadings >0.5 are given in bold
†Percentage of the variance explained by each factor and in total
EFA: Exploratory Factor Analysis, VDU: Visual Display Unit

Therefore, principal component analysis with oblimin rotation was repeated forcing two factors (table 3). The two-factor solution explained a total of 55,2% of the variance, with factor 1 contributing 38,9% and factor 2 contributing 16,3%. Examination of the factor loadings showed that the items “prolonged sitting” (2) and “prolonged visual display units (VDU) work” (3) loaded highly negative on the first factor and below 0.5 on the second factor and were therefore excluded. Item 22 “climbing stairs” was excluded due to negative loading on factor 2 and loading below 0.5 on factor 1. Item 21 “handling peddles with feet” showed low communalities value (0.276), indicating a poor fit with the other items in the factor and were therefore excluded. Item 7 “twisted posture” loaded above 0.5 on both factors, but slightly higher on factor 1 and were therefore retained in factor 1. As a result, 22 items remained (15 items in factor 1 and 7 items in factor 2).

Finally, a forced two-factor analysis with oblimin rotation on the remaining items was found to explain 58,3% of the total variance, with factor 1 contributing 41.6% and factor 2 contributing 16.7% (table 4). There was a weak correlation between the two factors (r=0.16) which indicates that the factors are representing different aspects of physical workload (22). The items that loaded high on the first factor were related to heavy physical work and items that loaded high on the second factor were related to static postures or repetitive movements. The factor labels proposed by Bot et al. in the original study suited the extracted factors in this analysis which resulted in naming the first factor: subscale 1 “Heavy physical workload” and the second factor: subscale 2 “Long lasting postures and repetitive movements”.

Table 3 Pattern and Structure matrix of two-factor solution after EFA with oblimin rotation

Item	**Pattern coefficient**		**Structure coefficient**		**Communalities**
Does your work involve	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**	
1. Standing	**0.68**	-0.13	**0.67**	-0.06	0.462
2. Sitting	-0.72	0.36	-0.68	0.28	0.589
3. VDU work	-0.73	0.39	-0.69	0.31	0.622
4. Walking	**0.74**	-0.13	**0.72**	-0.05	0.541
5. Kneeling/squatting	**0.71**	0.09	**0.72**	0.16	0.527
6. Repetitive movement	-0.83	**0.83**	0.00	**0.82**	0.674
7. Twisted posture	**0.57**	**0.51**	**0.63**	**0.57**	0.652
8. Neck bent forward	0.18	**0.63**	0.25	**0.65**	0.450
9. Turning/bending neck	0.32	**0.63**	0.39	**0.66**	0.539
10. Wrists bent or twisted	0.11	**0.74**	0.18	**0.75**	0.571
11. Hands above shoulders	**0.76**	0.19	**0.78**	0.27	0.640
12. Hands below knees	**0.77**	0.07	**0.78**	0.15	0.605
13. Moving loads (>5kg)	**0.81**	0.12	**0.82**	0.21	0.685
14. Moving loads (>25kg)	**0.75**	0.12	**0.76**	0.20	0.587
15. Exert force with arms	**0.74**	0.36	**0.78**	0.44	0.736
16. Maximal force exertions	**0.68**	0.35	**0.71**	0.42	0.630
17. Physical hard work	**0.82**	0.23	**0.84**	0.32	0.761
18. Static posture	-0.40	**0.80**	-0.31	**0.76**	0.735
19. Uncomfortable posture	0.37	**0.60**	0.43	**0.64**	0.537
20. Work with vibrating tools	**0.59**	0.07	**0.59**	0.13	0.356
21. Handling peddles with feet	**0.53**	-0.02	**0.53**	0.04	**0.276**
22. Climbing stairs	0.41	-0.15	0.39	-0.11	**0.174**
23. Often squatting	**0.76**	-0.16	**0.74**	-0.08	0.578
24. Walking on irregular surfaces	**0.62**	0.02	**0.63**	0.08	0.392
25. Sitting/moving on knees	**0.72**	0.03	**0.73**	0.11	0.530
26. Repetitive tasks arms/hands	-0.22	**0.71**	-0.14	**0.68**	0.511
Eigenvalue*	10.12	4.24			
Variance explained before rotation†	38.9%	16.3%			
Total variance explained†	55.2%				

Factor loadings >0.5 and communalities values <0.3 are given in bold
*Eigenvalues refer to the total variance explained by each factor
†Percentage of the variance explained by each factor and in total
EFA: Exploratory Factor Analysis, VDU: Visual Display Unit

Table 4 Final factor loadings after forced two factor solution and exclusion of items

Item	Pattern coefficients		Structure coefficients	
Does your work involve	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**
17. Physical hard work	**0.84**	0.15	**0.86**	0.29
13. Moving loads (>5kg)	**0.83**	0.04	**0.83**	0.17
12. Hands below knees	**0.80**	-0.02	**0.79**	0.11
11. Hands above shoulders	**0.78**	0.12	**0.80**	0.25
23. Often squatting	**0.78**	-0.25	**0.74**	-0.12
14. Moving loads (>25kg)	**0.77**	0.03	**0.77**	0.16
15. Exert force with arms	**0.75**	0.30	**0.80**	0.42
5. Kneeling/squatting	**0.75**	0.00	**0.75**	0.12
4. Walking	**0.73**	-0.18	**0.70**	-0.06
25. Sitting/moving on knees	**0.71**	-0.06	**0.74**	0.07
16. Maximal force exertions	**0.70**	0.29	**0.74**	0.40
1. Standing	**0.66**	-0.15	**0.64**	-0.04
24. Walking on irregular surfaces	**0.61**	-0.07	**0.64**	0.04
7. Twisted posture	**0.60**	0.46	**0.61**	**0.56**
20. Work with vibrating tools	**0.57**	0.04	**0.58**	0.13
6. Repetitive movement	-0.06	**0.83**	0.07	**0.82**
18. Static posture	-0.37	**0.82**	-0.24	**0.76**
10. Wrists bent or twisted	0.11	**0.75**	0.23	**0.77**
26. Repetitive tasks arms/hands	-0.20	**0.72**	-0.08	**0.69**
8. Neck bent forward	0.19	**0.63**	0.29	**0.66**
9. Turning/bending neck	0.33	**0.62**	0.43	**0.67**
19.Uncomfortable posture	0.39	**0.56**	0.49	**0.62**
Eigenvalue*	9.15	3.66		
Variance explained†	41.6%	16.7%		
Total variance explained†	58.3%			

Factor loadings >0.5 are given in bold
*Eigenvalues refer to the total variance explained by each factor
†Percentage of the variance explained by each factor and in total

Internal consistency

The Cronbach alpha value for both subscales were good and the item-total correlations indicated all items to correlate well with the total subscales. Values are shown in table 5.

Table 5 Internal consistency and item-total correlation of the subscales

Internal consistency (Cronbach's alpha)	
Subscale 1: Heavy physical workload	0.94
Subscale 2: Long lasting postures and repetitive movements	0.85
Item-total correlation	
Subscale 1: Heavy physical workload	0.53-0.84
Subscale 2: Long lasting postures and repetitive movements	0.52-0.73

Data quality

The proportion of missing data in the cross-sectional study was very small, under 4% for all items (ranging from 0,9% to 3,5%). For subscale 1 and 2, missing data was 7% and 4,3%, respectively. There were no floor or ceiling effects of the two subscales, but on the individual items, there were ceiling effect on five items and floor effect on all items, except from two items, with the highest floor effect of 79,8% (see additional file 2). The sum score on subscale 1 had a skewed distribution while subscale 2 was normally distributed. The distribution of difference score between test and retest were consistent with a normal distribution on both subscales, therefore the reliability analyses are presented with mean (SD) (22).

Construct validity

For six of the eight first hypotheses, the hypotheses that the median values were statistically significant different ($p<0.01$) among occupational groups were confirmed. For the hypotheses based on the other health-related constructs involved, all six hypotheses were confirmed. In total, 12 (85%) of the 14 predefined hypotheses were confirmed (table 6). Hence, more than 75% of the a priori hypotheses were achieved, indicating a good construct validity. Score per occupation group are shown in table 7. Although three of the occupation groups were normally distributed, the combination with the other group(s) had at least one group with skewed distribution, therefore non-parametric analyses were used (22) and all scores are presented with median (min-max).

Table 6 Construct validity: A priori formulated hypotheses (n=115)

	Value	Hypothesis confirmed
The mean score on the subscale “heavy physical workload” is significantly higher for the occupational group with heavy physical load (*n* = 10) than for the occupational group with long-lasting postures and repetitive movements (*n* = 52)	p = 0.008	Yes
The mean score on the subscale “heavy physical workload” is significantly higher for the occupational group with heavy physical load (*n* = 10) than for the occupational group without physical load (*n* = 20)	p = 0.038	No
The mean score on the subscale “long lasting postures and repetitive movements” is significantly higher for the occupational group with long lasting postures and repetitive movements (*n* = 53) than for the occupational group with heavy physical load (*n* = 10)	p = 0.001	Yes
The mean score on the subscale “long lasting postures and repetitive movements” is significantly higher for the occupational group with long lasting postures and repetitive movements (*n* = 53) than for the occupational group without physical load (*n* = 22)	p < 0.001	Yes
The mean score on the subscale “heavy physical workload” is significantly higher for the occupational group with both heavy physical load and long lasting postures and repetitive movements (*n* = 14) than for the occupational group without physical load (*n* = 20)	p < 0.001	Yes
The mean score on the subscale “long lasting postures and repetitive movements” is significantly higher for the occupational group with both heavy physical load and long lasting postures and repetitive movements (*n* = 13) than for the occupational group without physical load (*n* = 22)	p = 0.001	Yes
In the occupational group with heavy physical load (*n* = 10), the mean score on the subscale “heavy physical workload” is significantly higher than on the subscale “long lasting postures and repetitive movements”	p = 0.173	No
In the occupational group with long lasting postures and repetitive movements (*n* = 52), the mean score on the subscale “long lasting postures and repetitive movements” is significantly higher than on the subscale “heavy physical workload”	p < 0.001	Yes
High correlation between “heavy physical workload” score and the Mechanical Exposure Index score (*n* = 104)	rho = 0.783*	Yes
Moderate to high correlation between “long lasting postures and repetitive movements” score and the Mechanical Exposure Index score (*n* = 106)	rho = 0.402*	Yes
Low correlation between “heavy physical workload” score and physical functioning score (SF-36) (*n* = 107)	rho = -0.288	Yes
Low correlation between “long lasting postures and repetitive movements” score and physical functioning score (SF-36) (*n* = 110)	rho = -0.155	Yes
Low correlation between “heavy physical workload” score and general health score (SF-36) (*n* = 107)	rho = -0.126	Yes
Low correlation between “long lasting postures and repetitive movements” score and general health score (SF-36) (*n* = 110)	rho = -0.282	Yes

Hypotheses are analysed with Mann-Whitney U test, Wilcoxon signed ranks test (both presented with p values) and correlation analysis (presented with Spearman’s rho).
P values less than 0.01 were considered statistically significant.
Correlation coefficients under 0.3, between 0.3 and 0.6 and over 0.6 were considered low, moderate and high, respectively (27). *P values <0.01

Table 7 Scores per occupation group

	Heavy physical workload (range 0-100)		Long lasting postures and repetitive movements (range 0-100)	
	n		*n*	
No physical load	20‡	14.4 (0 - 82.2)	22	26.2 (0 - 100)
Heavy physical load	10†	32.2 (0 - 64.4)	10†	23.8 (0 - 71.4)
Long lasting postures and repetitive movements	52‡	4.4 (0 - 66.7)	53†	57,1 (4,8 - 100)
Both heavy physical load and long lasting postures and repetitive movements	14	55.6 (28.9 - 77.8)	13†	66.7 (4.8 - 95.2)
Missing*	14		14	

† No subscale score could be calculated for one subject of the group due to incomplete questionnaire

‡No subscale score could be calculated for two subjects of the group due to incomplete questionnaires

*Occupations that were unknown or could not be classified.

Reliability and measurement error

Relative and absolute reliability values for participants reporting no change in physical workload are presented in table 8. Both subscales showed acceptable relative reliability ($ICC_{2,1}$ >0.7). Sensitivity analysis which included five participants who reported change in the physical workload from first test to retest revealed similar results in all reliability analysis.

Table 8 Reliability and measurement error

	First test	Retest	Difference	$ICC_{2.1}$ (95% CI)	$SEM_{agreement}$	$SDC_{95\% ind}$
Heavy physical workload *(n=48)*	27.3 ± 24.7	22.2 ± 21.9	-5.1 ± 8.5*	0.96 (0.88, 0.98)	6.9	19.2
Long lasting postures and repetitive movements *(n=48)*	49.2 ± 25.4	42.7 ± 24.1	-6.5 ± 12.6*	0.92 (0.81, 0.96)	10.0	27.7

ICC: intraclass correlation coefficient, SEM: standard error of measurement, SDC: smallest detectable change. Score values and difference values are presented with mean ±SD.

*P values <0.01

Discussion

Factor analysis revealed that the PWQ could be divided into two subscales. Both subscales showed good internal consistency with Cronbach's alpha of 0.94 and 0.85, on subscale 1 and 2, respectively. The current study also assessed validity and reliability of the Norwegian version of the PWQ established in a group of patients with various musculoskeletal disorders. The validity was considered acceptable as >75% of the predefined hypotheses were confirmed and the reliability was considered good as both subscales were above the recommended ICC of 0.70 (14). These results support the dimensionality of the questionnaire and suggest that the PWQ is a valid and reliable measure of physical workload in a Norwegian population with various musculoskeletal disorders.

Since the questions drawn from the DMQ, forming the basis for the development of PWQ, only have been examined once previously (in the original study of Bot et al.), all 26 questions extracted from DMQ were translated into Norwegian and included in the factor analyses. A further argument to include all 26 questions was that the BACE study targets patients with back pain, while the population included in the original study mainly had upper and lower extremity pain. As a result of the EFA, two subscales with a total of 22 items remained in the final PWQ. The present results were in line with the original study of Bot et al. in terms of the number of subscales obtained and also in terms of the nature of the items comprising each of the subscales, which could be labelled; subscale 1 "Heavy physical workload" and subscale 2 "Long lasting postures and repetitive movements". However, number of items included in each subscale differed. In the current study, 3 more items were included in subscale 1; "work with vibrating tools", "walking on irregular surfaces" and "twisted posture", and one more item were included in subscale 2: "uncomfortable posture". This difference was not entirely unexpected in light of the difference of the participant characteristics in the two studies. The most frequently reported pain area in the present study was the back region. Back pain is in several previous studies found to be associated with risk factors such as twisted posture (28-30), working with vibrating tools (31) and uncomfortable postures/ awkward postures (28, 29, 32). This might explain why these items loaded strong enough to be included in the subscales in the present study. Also the item "walking on irregular surfaces" may be associated with participants with low back pain, as previous studies have indicated that uneven surfaces are more challenging for postural control which can cause problems when handling loads and requires the worker to compensate by placing the body in awkward positions or to put more force into the task (33-35). Although parallel analysis supported the two- factor solution in the present study, the three-factor solution was assessed to make sure we did the right decision regarding extraction of the factors. Since the three-factor solution revealed only two items to be obtained in the third factor and as a subscale in a questionnaire should be comprised of least three items (26), we went further with the two-factor solution. However, the two items in the third factor in the three-factor solution: "sitting" and "prolonged visual display units work" are indicating sedentary work, which is common nowadays, and are in several studies shown to be associated with mortality and chronic health disorders (36, 37). Therefore, it would be interesting to make a third subscale, "sedentary work", although it would require to add extra items formulated in the same aspect, which further would require a new field study (26).

The good internal consistency of the subscales, 0.94 for the "heavy physical workload" and 0.85 for the "long-lasting postures and repetitive movements" indicates that the items in the respectively subscales correlates well with each other, thus measuring the same concept (14). However, an alpha exceeding 0.9, may indicate that some items are redundant, which means they are testing the same question but in a different way (15). Examination of the item-total statistics shows that two items would decrease the Cronbach's alpha to 0.93 if they were removed from the scale ("moving loads >5kg" and "physical hard work"), however the gap is so minimal and we consider the items to be important and to contribute to the content validity of the instrument. The Cronbach's alpha coefficient can be sensitive to the number of items in a scale, and questionnaires or subscales with fewer than 10 items can result in a value that is too low (22, 26). Despite this, the "long lasting postures and repetitive movements" subscale which consist of only 7 items, had a Cronbach's alpha of 0.85. Item-total correlation ranged from 0.53 to 0.84 in subscale 1, and from 0.52 to 0.73 in subscale 2, indicating all items to correlate well with the total subscales (26). This result is consistent with the original study of Bot et al., showing Cronbach`s alpha of 0.92-0.93 on the "heavy physical workload" subscale and 0.86-0.87 on the "long lasting postures and repetitive movements subscale" (from three different study populations with upper or lower extremity complaint).

Construct validity is an important element of the validity of a questionnaire. For six of the eight first hypotheses regarding "known" group validity, there was enough evidence to confirm the hypotheses that the median values were statistically significant different ($p<0.01$) among occupational groups. There is a large interindividual variability within jobs, and as there were several disagreements whether to classify jobs in the "heavy physical load" group or in the group with "both heavy physical load and long lasting postures and repetitive movements", we may have misclassified a substantial number of subjects. As the present study revealed different items than Bot et al. to be included in the two subscales, we must be careful with comparing our results with the original study. However, there are a few similarities to point out anyway; in line with Bot et al., we found that the PWQ could clearly distinguish between the scores of the occupational group with "long lasting postures and repetitive movements" as this group scored low on the first subscale and high on the second. In addition, all hypotheses regarding occupations classified as "both physical heavy load and long lasting postures and repetitive movements" were confirmed, which further provides initial evidence for the validity of the questionnaire. The hypotheses regarding convergent validity and discriminating validity were confirmed. The MEI, which is a self-report questionnaire assessing mechanical exposure of the shoulder- neck region (21), shows high correlation with the “heavy physical workload” subscale and moderate correlation with the “long lasting postures and repetitive movements” subscale. The SF-36 dimensions, “physical function” and general health (20), measure other constructs than physical workload, and as expected, low correlation was found with both subscales. The validity analyses confirmed 85% of the predefined hypotheses, indicating a good construct validity (14).

The ICC was well above the minimum standard of both subscales, 0.96 and 0.92 on subscale 1 and 2, respectively, and therefore considered to be acceptable, which suggest that the PWQ is a reliable measure in our population (26). The absolute reliability, presented as measurement error and reported in the actual scale unit, is more clinically useful than the relative reliability. The $SDC_{95\%ind}$ of 19.2 and 27.7 on subscale 1 and 2, respectively, indicates that a score of self- reported physical workload at the individual level would have to change by 19.2 and 27.7 on subscale 1 and 2, respectively, to ensure that the change was not a result of measurement error (26). Since we were unable to identify other studies assessing the measurement properties of the PWQ, and Bot et al. only evaluated internal consistency and validity, we believe this is the first study to determine the test-retest reliability of the questionnaire, and we have no other studies to compare with. To determine whether the SEM and SDC values are acceptable depends on which changes are minimally important on the PWQ subscales. The SDC should be smaller than the minimal important change (MIC) (14), however, no MIC for the PWQ scales are available, and we were unable to determine whether the SDC of the PWQ scales were sufficiently low. However, these relatively high SEM and SDC values are similar to quality of life outcomes on the SF-36 scale among patients with chronic neck pain (38), and the SF-36 has widely been used for its good measurement properties. Despite this, on a scale from 0-100, these values may indicate relatively large measurement error, and it must be taken into consideration by researchers and clinicians who use the scale on the individual level in clinical practice to evaluate change over time. Further research is needed to evaluate the responsiveness (sensitivity to change) and MIC of the PWQ subscales.

The COSMIN group recommends a minimum of 50 participants in test-retest analyses. Although 62 participated in the test-retest study, we excluded 14 participants due to change in the working condition from test to retest or due to incomplete questionnaires. Therefore, only 48 participants were included in the test- retest analyses and there might be some imprecision in our estimates regarding test-retest assessment. The individuals participating in the re-test reported more back pain but less pain in neck and shoulder/upper extremity, and they also reported slightly better general health and lower physical function on SF-36. Although the patients participating in the test and the re-test were similar, these dissimilarities may have affected the results. Test-retest reliability should be assessed in a stable population with an appropriate time interval between measurements (14). In the current study the time interval was median 3 days (range 1-10), which means shorter than recommended for many of the participants. There is a potential risk of recall bias if the interval between the test and the retest is too short. We believe though, that the comprehensive questionnaire with a high number of questions filled out at the first test most likely reduce recall bias when filled out only few days later. Regarding stable population, five participants reported change in the physical workload from first test to retest and were therefore excluded from the reliability analysis. However, sensitivity analysis revealed similar results in all reliability analysis indicating the results to be reliable.

Strengths and limitations

One strength of this study is that we reduced the number of the 26 items in a systematic manner by performing an EFA according to guidelines. Another strength is that the assessment was following the COSMIN checklist and the PROM guidelines (13, 14). We had more than 100 participants in the validity study, and nearly 50 in the reliability part. Furthermore, in addition to “known” group validity, the construct validity was assessed by convergent and discriminating validity. Lastly, this study is the first to assess the reliability of the PWQ.

A limitation of this study is that the item pool originally was chosen because the items were expected to have an association with either upper- or lower extremity complaints. This actually raises the question of whether the questionnaire can be used to assess physical workload in any other population than patients with upper- or lower extremity complaints as there might be dimensions not taken into account to cover all aspects of musculoskeletal disorders. As the participants in the present study had various musculoskeletal pain, this may have affected the results. Another limitation is the use of self- rating in the assessment. Although self-reported questionnaires are known for being sufficient to rank the physical workload of specific activities, tasks or jobs (39), self-ratings may suffer from misclassification. There are studies showing that workers with musculoskeletal disorders may overestimate the physical load compared to healthy workers (40, 41). Even when participants are motivated to report the workload accurately, they may have difficulties with recalling and accurately reporting the information, which may threaten the validity of the questionnaire (42). Furthermore, several aspects of the sample may have affected the results. Our sample are recruited from a clinic located in Asker, a wealthy town close to the capital of Norway, which may imply that this study’s population consist of participants with high socioeconomic status. Previous studies showed that low socioeconomic status is associated with higher exposure of physical workload (43, 44) and there is reason to believe that by recruiting participants from a wider geographical area we would have reached a broader population regarding occupational variation. In this study more women than men were represented, which could threaten the generalizability of the results. However, musculoskeletal disorders are shown to be more prevalent in women (45), and also that women are more prevalent patients in outpatient clinics (46). Therefore, the gender distribution of the present study is probably representative of the patients seeking treatment for musculoskeletal disorders in outpatient physiotherapy clinics. There were 14 participants who were not in paid job (occupational rehabilitation, student, homemakers, disabled and unemployed). They were included in the analyses as they responded to the questionnaire, and we think that physical workload can occur regardless if the job is paid or not, however they may have affected the results. We also have a lack of data on eligible study participants that declined to participate. Due to limited resources it was not possible to record information on all patients attending the rehabilitation clinic during the data collection period. We must also acknowledge that cultural differences can affect measurement properties of a questionnaire (47), and that the translation of PWQ into Norwegian may not have succeeded in establishing the exact same meaning for each and every item, though the recommendations for cultural adaptations of questionnaires were followed.

Conclusions

The Norwegian version of PWQ, consisting of two subscales: “Heavy physical workload” (15 items) and “Long lasting postures and repetitive movements” (7 items), shows acceptable internal consistency, construct validity and reliability when used on patients with various musculoskeletal disorders receiving rehabilitation in a secondary outpatient clinic in Norway. The self-administered questionnaire is easy and cheap to use in clinical practice and research. This study indicates that the PWQ can be used in clinical and occupational healthcare and in research purposes on patients with musculoskeletal disorders. However, further research is advised on test-retest reliability and construct validity in other populations and clinical settings. Also, the clinical value of the PWQ in relation to work-related musculoskeletal disorders should be investigated.

List of abbreviations

BACE: BAck Complaints in Elders

CI: Confidence Interval

COSMIN: COnsensus-based Standards for the selection of health Measurement INstruments

DMQ: Dutch Musculoskeletal Questionnaire

EFA: Exploratory Factor Analysis

ICC: Intraclass Correlation Coefficient

iPCQ: iProductivity Cost Questionnaire

MEI: Mechanical Exposure Index

MIC: Minimal Important Change

NRS: Numeric Rating Scale

NSD: Norwegian Centre for Research Data

PCA: Principal Component Analysis

PROMs: Patient-Related Outcome-Measures

PWQ: Physical Workload Questionnaire

QPSnordic: General Nordic Questionnaire for Psychological and social factors at work

REC: Regional Ethical Committees for Medicine and Health

SD: Standard Deviation

SDC: Smallest Detectable Change

SEM: Standard Error of Measurement

SF-36: Short Form 36 Health Status Questionnaire

VDU: Visual Display Unit

Declarations

Ethics approval and consent to participate

The study is considered a quality assessment project by the Regional Ethical Committees for Medicine and Health (REC) (reference no. 2014/1634/REK vest). Approval for this study was obtained from the Norwegian Centre for Research Data (NSD) (reference No. 42149) in 2019. All participants signed a written informed consent prior to inclusion in the study.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and analysed during the study are available from the project manager of BACE-N, Margreth Grotle, on reasonable request.

Competing interests

The author declares to have no competing interests.

Funding

This study was supported by Oslo Metropolitan University.

Authors' contributions

Not applicable

Acknowledgements

Ingeborg Landstad and Unicare Friskvern and the BACE-N team contributed to data collection.

References

1. Da Costa BR, Vieira ER. Risk factors for work-related musculoskeletal disorders: a systematic review of recent longitudinal studies. American journal of industrial medicine. 2010;53:285-323.

2. Mayer J, Kraus T, Ochsmann E. Longitudinal evidence for the association between work-related physical exposures and neck and/or shoulder complaints: a systematic review. International Archives of Occupational and Environmental Health. 2012;85(6):587-603.

3. Andersen LL, Fallentin N, Thorsen SV, Holtermann A. Physical workload and risk of long-term sickness absence in the general working population and among blue-collar workers: prospective cohort study with register follow-up. Occupational & Environmental Medicine. 2016;73(4):246-53.

4. Sterud T. Work-related mechanical risk factors for long-term sick leave: a prospective study of the general working population in Norway. The European Journal of Public Health. 2014;24(1):111-6.

5. Bang CK, Lund T, Labriola M, Villadsen E, Bültmann U. The fraction of long-term sickness absence attributable to work environmental factors: prospective results from the Danish Work Environment Cohort Study. Occupational & Environmental Medicine. 2007;64:487-9.

6. Griffith LE, Shannon HS, Wells RP, Walter SD, Cole DC, Côté P, et al. Individual participant data meta-analysis of mechanical workplace risk factors and low back pain. American journal of public health. 2012;102(2):309-18.

7. Stock SR, Fernandes R, Delisle A, Vezina N. Reproducibility and validity of workers' self-reports of physical work demands. Scandinavian Journal of Work, Environment & Health. 2005;31(6):409-37.

8. David GC. Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders. Occupational Medicine. 2005;55(3):190-9.

9. van Der Beek AJ, Frings-Dresen MH. Assessment of mechanical exposure in ergonomic epidemiology. Occupational & Environmental Medicine. 1998;55(5):291-9.

10. Barrero LH, Katz JN, Dennerlein JT. Validity of self-reported mechanical demands for occupational epidemiologic research of musculoskeletal disorders. Scandinavian Journal of Work, Environment & Health. 2009;35(4):245-60.

11. Bot SDM, Terwee C, van Der Windt DAWM, Feleus A, Bierma-Zeinstra SM, Knol DL, et al. Internal consistency and validity of a new physical workload questionnaire. Occupational & Environmental Medicine. 2004;61:980-6.

12. Hildebrandt VH, Bongers PM, van Dijk FJH, Kemper HCG, Dul J. Dutch Musculoskeletal Questionnaire: description and basic qualities. Ergonomics. 2001;44(12):1038-55.

13. Mokkink L, Terwee C, Patrick D, Alonso J, Stratford P, Knol D, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research. 2010;19(4):539-49.

14. Terwee CB, Bot SDM, de Boer MR, van Der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. 2007;60(1):34-42.

15. Nunnally JC, Bernstein IH. Psychometric theory: McGraw-Hill; 1994.

16. Öhlund C, Eek C, Palmblad S, Areskoug B, Nachemson A. Quantified Pain Drawing in Subacute Low Back Pain: Validation in a Nonselected Outpatient Industrial Sample. Spine. 1996;21(9):1021-30.

17. Von Korff M, Jensen MP, Karoly P. Assessing Global Pain Severity by Self-Report in Clinical and Health Services Research. Spine. 2000;25(24):3140-51.

18. Dallner M. Validation of the General Nordic Questionnaire (QPSNordic) for psychological and social factors at work. København: Nordisk Ministerråd; 2000.

19. Bouwmans C, Krol M, Severens H, Koopmanschap M, Brouwer W, Roijen LV-H. The iMTA Productivity Cost Questionnaire: A Standardized Instrument for Measuring and Valuing Health-Related Productivity Losses. Value in Health. 2015;18(6):753-8.

20. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Medical Care. 1992;30(6):473-83.

21. Balogh I, Ørbaek P, Winkel J, Nordander C, Ohlsson K, Ektor-Andersen J. Questionnaire-based mechanical exposure indices for large population-reliability, internal consistency and predictive validity. Scandinavian Journal of Work, Environment and Health. 2001;27(1):41-8.

22. Pallant J. SPSS survival manual: a step by step guide to data analysis using IBM SPSS. 6 ed. Maidenhead: McGraw Hill; 2016.

23. Tabachnick BG, Fidell LS. Using multivariate statistics. 6 ed. Harlow: Pearson; 2014.

24. Cattell RB. The Scree Test For The Number Of Factors. Multivariate Behavioral Research. 1966;1(2):245-76.

25. Watkins M. Monte Carlo PCA for Parallel Analysis (Computer Software). State College, PA: Ed & Psych Associates. 2000.

26. de Vet H, Terwee C, Mokkink L, Knol D. Measurement in Medicine: A Practical Guide. Cambidge: Cambridge University Press; 2011.

27. Andresen EM. Criteria for assessing the tools of disability outcomes research. Archives of Physical Medicine and Rehabilitation. 2000;81:15-20.

28. Videman T, Ojajarvi A, Riihimaki H, Troup JD. Low back pain among nurses: a follow-up beginning at entry to the nursing school. Spine. 2005;30(20):2334-41.

29. Van Nieuwenhuyse A, Somville PR, Crombez G, Burdorf A, Verbeke G, Johannik K, et al. The role of physical workload and pain related fear in the development of low back pain in young workers: evidence from the BelCoBack Study; results after one year of follow up. Occupational & Environmental Medicine. 2006;63(1):45-52.

30. Campo M, Weiser S, Koenig KL, Nordin M. Work-related musculoskeletal disorders in physical therapists: a prospective cohort study with 1-year follow-up. Phys Ther. 2008;88(5):608-19.

31. Bovenzi M, Schust M, Mauro M. An overview of low back pain and occupational exposures to whole-body vibration and mechanical shocks. La Medicina del lavoro. 2017;108:419-33.

32. Bovenzi M. Metrics of whole-body vibration and exposure-response relationship for low back pain in professional drivers: a prospective cohort study. Int Arch Occup Environ Health. 2009;82(7):893-917.

33. Gates DH, Scott SJ, Wilken JM, Dingwell JB. Frontal plane dynamic margins of stability in individuals with and without transtibial amputation walking on a loose rock surface. Gait & Posture. 2013;38(4):570-5.

34. Marigold DS, Patla AE. Age-related changes in gait for multi-surface terrain. Gait & Posture. 2008;27(4):689-96.

35. Thies SB, Richardson JK, Ashton-Miller JA. Effects of surface irregularity and lighting on step variability during gait: A study in healthy young and older women. Gait & Posture. 2005;22(1):26-31.

36. Diaz KM, Howard VJ, Hutto B, Colabianchi N, Vena JE, Safford MM, et al. Patterns of Sedentary Behavior and Mortality in U.S. Middle-Aged and Older Adults: A National Cohort Study. Ann Intern Med. 2017;167(7):465-75.

37. Wilmot E, Edwardson C, Achana F, Davies M, Gorely T, Gray L, et al. Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis. Clinical and Experimental Diabetes and Metabolism. 2012;55(11):2895-905.

38. Juul T, Søgaard K, Davis AM, Roos EM. Psychometric properties of the Neck OutcOme Score, Neck Disability Index, and Short Form–36 were evaluated in patients with neck pain. Journal of Clinical Epidemiology. 2016;79:31-40.

39. Burdorf A, Van Der Beek AJ. In musculoskeletal epidemiology are we asking the unanswerable in questionnaires on physical load? Scandinavian Journal of Work, Environment & Health. 1999;25(2):81-3.

40. Balogh I, Ørbæk P, Ohlsson K, Nordander C, Unge J, Winkel J, et al. Self-assessed and directly measured occupational physical activities—influence of musculoskeletal complaints, age and gender. Applied Ergonomics. 2004;35(1):49-56.

41. Hansson G-Å, Balogh I, Byström J, Ohlsson K, Nordander C, Asterland P, et al. Questionnaire versus direct technical measurements in assessing postures and movements of the head, upper back, arms and hands. Scandinavian Journal of Work, Environment & Health. 2001(1):30-40.

42. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc. 2016;9:211-7.

43. Aittomäki A, Lahelma E, Roos E, Leino-Arjas P, Martikainen P. Gender Differences in the Association of Age with Physical Workload and Functioning. Occupational & Environmental Medicine. 2005;62(2):95-100.

44. Mehlum IS, Kristensen P, Kjuus H, Wergeland E. Are Occupational Factors Important Determinants of Socioeconomic Inequalities in Musculoskeletal Pain? Scandinavian Journal Of Work, Environment & Health. 2008;34(4):250-9.

45. Ihlebæk C, Brage S, Natvig B, Bruusgaard D. Forekomst av muskel- og skjelettlidelser i Norge. Tidsskrift for Den norske legeforening. 2010;130(23):2365-8.

46. Kinge JM, Knudsen AK, Skirbekk V, Vollset SE. Musculoskeletal disorders in Norway: prevalence of chronicity and use of primary and specialist health care services. BMC Musculoskeletal Disorders. 2015;16(1).

47. Streiner DL, Norman GR, Cairney J. Health Measurement Scales: A practical guide to their development and use. Fifth edition. ed: United Kingdom: Oxford University Press; 2014.

Additional files

Additional file 1. The Norwegian version of the PWQ

Spørreskjema for fysisk arbeidsbelastning

Innebærer arbeidet ditt…		**sjelden eller aldri**	**av og til**	**ofte**	**(nesten) alltid**
1.	lange perioder hvor du står?	1	2	3	4
2.	lange perioder hvor du sitter?	1	2	3	4
3.	lange perioder med skjermarbeid?	1	2	3	4
4.	lange perioder hvor du går?	1	2	3	4
5.	lange perioder hvor du står på kne eller på huk?	1	2	3	4
6.	lange perioder hvor du gjentar de samme bevegelsene?	1	2	3	4
7.	lange perioder hvor du arbeide i en vridd stilling?	1	2	3	4
8.	lange perioder hvor du er i en framoverbøyd eller vridd stilling med nakken?	1	2	3	4
9.	at du ofte må bøye eller vri hodet?	1	2	3	4
10.	lange perioder hvor du har en bøyd eller vridd stilling i håndleddet?	1	2	3	4
11.	at du må arbeide med hendene over skulderhøyde?	1	2	3	4
12.	at du må arbeide med hendene under knehøyde?	1	2	3	4
13.	at du må forflytte objekter over 5 kg?	1	2	3	4
14.	at du må forflytte tunge objekter over 25 kg?	1	2	3	4
15.	at du bruker krefter i armer eller hender?	1	2	3	4
16.	at du tar i alt du kan?	1	2	3	4
17.	tungt fysisk arbeid?	1	2	3	4

18.	at du arbeider i den samme stillingen i lange perioder?	1	2	3	4
19.	at du arbeider i ukomfortable stillinger?	1	2	3	4
20.	arbeid med vibrerende verktøy/instrumenter?	1	2	3	4
21.	at du bruker eller styrer pedaler med føttene?	1	2	3	4
22.	at du går i trapper?	1	2	3	4
23.	at du ofte går opp og ned på huk	1	2	3	4
24.	at du går på ujevnt underlag?	1	2	3	4
25.	at du sitter eller beveger deg på knærne?	1	2	3	4
26.	repeterende arbeidsoppgaver med armene, hendene eller fingrene mange ganger i minuttet?	1	2	3	4

Physical workload questionnaire etter Bot SD et al. Occup Environ Med. 2004 Dec;61(12):980-6:
Internal consistency and validity of a new physical workload questionnaire. Oversatt av Grotle M og Munk R 2014, HiOA.

Additional file 2 Table of missing data, floor- and ceiling effects of the PWQ

Missing data, floor- and ceiling effects for the PWQ subscales and items (n=115)

PWQ items. Does your work involve…	**Missing data, n (%)**	**Lowest (%)**	**Highest (%)**
Heavy physical workload	**8 (7.0)**	**9.3**	**0.0**
1. Standing for long periods of time?	1 (0.9)	2.2	20.2
4. Walking long periods of time?	1 (0.9)	42.1	8.8
5. Kneeling or squatting for long periods of time?	2 (1.7)	64.6	3.5
7. Working in a twisted posture for long periods of time?	1 (0.9)	48.2	5.3
11. Work(ing) with your hands above shoulder level?	1 (0.9)	58.8	3.5
12. Work(ing) with your hands below knee level?	1 (0.9)	67.5	0.0
13. Moving loads (more than 5 kg)?	1 (0.9)	49.1	11.4
14. Moving heavy loads (more than 25 kg)?	2 (1.7)	63.7	0.9
15. Exerting force with your arms or hands?	2 (1.7)	42.5	14.2
16. Exerting maximal force?	1 (0.9)	59.6	7.9
17. Physical hard work?	2 (1.7)	59.3	6.2
20. Working with vibrating tools?	1 (0.9)	79.8	2.6
23. Squatting often?	2 (1.7)	52.2	6.2
24. Walking on irregular surfaces?	1 (0.9)	68.4	4.4
25. Sitting or moving on your knees?	4 (3.5)	72.2	1.8
Long lasting postures and repetitive movements	**5 (4.3)**	**1.8**	**4.5**
6. Making the same movement for long periods of time?	1 (0.9)	14.9	34.2
8. Holding your neck in a bent forward or twisted position for long periods of time?	1 (0.9)	30.7	13.2
9. Bending or twisting your neck often?	2 (1.7)	27.4	11.5
10. Holding your wrist in a bent or twisted position for long periods of time?	2 (1.7)	29.2	16.8
18. Working in the same position for long periods of time?	1 (0.9)	14.9	38.6
19. Working in uncomfortable postures?	3 (2.6)	32.1	8.9
26. Doing repetitive tasks with arms, hands or fingers many times per minute?	1 (0.9)	29.8	34.2

APPENDICES

Appendix 1 Submission guidelines for BMC Public Health

Below are the main submission guidelines for submitting a research article in the BMC Public Health. The complete submission guidelines can be found at:

https://bmcpublichealth.biomedcentral.com/submission-guidelines/preparing-your-manuscript/research-article

The information below details the section headings that you should include in your manuscript and what information should be within each section. Please note that your manuscript must include a 'Declarations' section including all of the subheadings (please see below for more information).

Title page

The title page should:

- present a title that includes, if appropriate, the study design e.g.:
  - "A versus B in the treatment of C: a randomized controlled trial", "X is a risk factor for Y: a case control study", "What is the impact of factor X on subject Y: A systematic review"
  - or for non-clinical or non-research studies a description of what the article reports
- list the full names and institutional addresses for all authors
  - if a collaboration group should be listed as an author, please list the Group name as an author. If you would like the names of the individual members of the Group to be searchable through their individual PubMed records, please include this information in the "Acknowledgements" section in accordance with the instructions below
- indicate the corresponding author

Abstract

The Abstract should not exceed 350 words. Please minimize the use of abbreviations and do not cite references in the abstract. Reports of randomized controlled trials should follow the CONSORT extension for abstracts. The abstract must include the following separate sections:

- **Background:** the context and purpose of the study
- **Methods:** how the study was performed and statistical tests used
- **Results:** the main findings
- **Conclusions:** brief summary and potential implications
- **Trial registration:** If your article reports the results of a health care intervention on human participants, it must be registered in an appropriate registry and the registration number and date of registration should be in stated in this section. If it was not registered prospectively (before enrollment of the first participant), you should include the words 'retrospectively registered'. See our editorial policies for more information on trial registration

Keywords

Three to ten keywords representing the main content of the article.

Background

The Background section should explain the background to the study, its aims, a summary of the existing literature and why this study was necessary or its contribution to the field.

Methods

The methods section should include:

- the aim, design and setting of the study
- the characteristics of participants or description of materials
- a clear description of all processes, interventions and comparisons. Generic drug names should generally be used. When proprietary brands are used in research, include the brand names in parentheses
- the type of statistical analysis used, including a power calculation if appropriate

Results

This should include the findings of the study including, if appropriate, results of statistical analysis which must be included either in the text or as tables and figures.

Discussion

This section should discuss the implications of the findings in context of existing research and highlight limitations of the study.

Conclusions

This should state clearly the main conclusions and provide an explanation of the importance and relevance of the study reported.

List of abbreviations

If abbreviations are used in the text they should be defined in the text at first use, and a list of abbreviations should be provided.

Declarations

All manuscripts must contain the following sections under the heading 'Declarations':

- Ethics approval and consent to participate
- Consent for publication
- Availability of data and materials
- Competing interests
- Funding
- Authors' contributions
- Acknowledgements
- Authors' information (optional)

Appendix 2 Translation and cross-cultural adaptation procedure of the PWQ

The translation and cross-cultural adaptation were done according to international guidelines (Beaton, Bombardier, Guillemin, & Ferraz, 2000; Guillemin, Bombardier, & Beaton, 1993). As we wanted to construct a Norwegian version of the PWQ, not only verify the two subscales created by Bot et al., all the 26 items from the item pool were translated. Two translators (one philologist and one clinician), whose mother tongue is Norwegian, independently translated the 26 items into Norwegian and synthesized them into one Norwegian version before it was translated back to English. Two translators and native English speakers, blinded the original PWQ item pool, independently performed the backtranslation and synthesized the two versions into one English version. An expert committee consisting of the translators and two researchers from the research group reviewed all translations and agreed on a prefinal version. The goal was, that the prefinal Norwegian PWQ item pool should be as concise and easy to understand as possible. Ten patients with musculoskeletal disorders reviewed the prefinal Norwegian version. The items and responses were confirmed to be relevant and understandable without any proposed alterations. Since the prefinal version was acceptable and easy to comprehend, no changes were made and the final version of the Norwegian PWQ item pool was equal to the prefinal version.

Appendix 3 Information about participation and consent form

Forespørsel om deltakelse i forskningsprosjektet

Validering av spørreskjemaene PWQ og iPCQ «BACE: en prospektiv kohort studie av eldre med ryggsmerter som søker hjelp i primærhelsetjenesten»

Bakgrunn og hensikt

Dette er et spørsmål til deg om å delta i en forskningsstudie for å teste gyldigheten av to spørreskjemaer som nylig har blitt oversatt til norsk;. Oversettelsesarbeidet er en del av et større ryggstudie som har til formål å undersøke forløpet av ryggsmerter hos eldre. Fakultet for helsefag ved Høgskolen i Oslo og Akershus er hovedansvarlig for prosjektet.

Hva innebærer studien?

Studien innebærer at du to ganger må svare på et spørreskjema om helse og funksjon, fysisk arbeidsbelastning på jobb, redusert arbeidskapasitet og eventuell sykefravær.

Du vil få utlevert spørreskjemaene, med ca. 2 til 3 dagers mellomrom, på Unicare Friskvernklinikken.

Mulige fordeler og ulemper

Erfaring fra studien vil kunne bidra til at disse spørreskjemaene kan implementeres som et måleredskap for fysisk arbeidsbelastning på jobb og produktivitetskostnader ved helserelaterte plager, noe som vil være av betydning for å bedre forskningen som gjøres på muskelskjelettplager i Norge. De to spørreskjemaene er spesifikke for forskningsformål og forventes ikke å føre til ubehag for deltagerne i studien.

Hva skjer med informasjonen om deg?

Informasjonen som registreres om deg skal kun brukes slik som beskrevet i hensikten med studien. All informasjon blir lagret på en sikker forskningsserver ved IT-avdelingen, HiOA. Direkte personidentifiserende opplysninger erstattes med et referansenummer som viser til en atskilt navneliste (koblingsnøkkel). Navnelisten/koblingsnøkkelen vil oppbevares av prosjektleder. Kun prosjektleder og forskningskoordinator vil ha tilgang til navnelisten/koblingsnøkkelen.

Resultater fra prosjektet vil formidles i form av artikler publisert i nasjonale og internasjonale tidsskrift, samt deler av materialet vil benyttes i masteroppgaver ved HiOA. Resultatene vil også presenteres ved norske og internasjonale konferanser. Forskning knyttet til dette prosjektet vil skje i samsvar med Helseforskningsloven. Prosjektperioden starter etter at alle godkjennelser foreligger og dataene for prosjektet vil bestå til februar 2025. Etter 1.mars 2025 vil alle direkte og indirekte personopplysninger i prosjektet bli anonymisert eller slettet.

Frivillig deltakelse
Det er frivillig å delta i studien. Du kan når som helst og uten å oppgi noen grunn trekke ditt samtykke til å delta i studien. Dette vil ikke få konsekvenser for din videre behandling. Dersom du ønsker å delta, undertegner du samtykkeerklæringen på neste side. Om du nå sier ja til å delta, kan du senere trekke tilbake ditt samtykke uten at det påvirker din øvrige behandling. Dersom du senere ønsker å trekke deg eller har spørsmål til studien, kan du kontakte prosjektleder Margreth Grotle, tlf. 90111172.

Samtykke til deltakelse i forskningsprosjektet

Validering av spørreskjemaene PWQ og iPCQ «BACE: en prospektiv kohort studie av eldre med ryggsmerter som søker hjelp i primærhelsetjenesten»

Navn		ID nummer:
Adresse		
Telefonnummer		
Epostadresse		

1) Jeg er villig til å delta i studien

(Dato, signert av prosjektdeltaker)

2) Jeg bekrefter å ha gitt informasjon om studien

(Dato, signert, rolle i studien)

HØGSKOLEN I OSLO OG AKERSHUS