Review

Scand J Work Environ Health. 2012;38(1):5–18. doi:10.5271/sjweh.3190

Evaluation of the measurement properties of self-reported health-related work-functioning instruments among workers with common mental disorders

by Femke I Abma, MSc,[1] Jac JL van der Klink, PhD,[1] Caroline B Terwee, PhD,[2] Benjamin C III Amick, PhD,[3, 4] Ute Bültmann, PhD [1]

Abma FI, van der Klink JJL, Terwee CB, Amick BC III, Bültmann U. Evaluation of the measurement properties of self-reported health-related work-functioning instruments among workers with common mental disorders. *Scand J Work Environ Health.* 2012;38(1):5–18. doi:10.5271/sjweh.3190

Objectives During the past decade, common mental disorders (CMD) have emerged as a major public and occupational health problem in many countries. Several instruments have been developed to measure the influence of health on functioning at work. To select appropriate instruments for use in occupational health practice and research, the measurement properties (eg, reliability, validity, responsiveness) must be evaluated. The objective of this study is to appraise critically and compare the measurement properties of self-reported health-related work-functioning instruments among workers with CMD.

Methods A systematic review was performed searching three electronic databases. Papers were included that: (i) mainly focused on the development and/or evaluation of the measurement properties of a self-reported health-related work-functioning instrument; (ii) were conducted in a CMD population; and (iii) were fulltext original papers. Quality appraisal was performed using the consensus-based standards for the selection of health status measurement instruments (COSMIN) checklist.

Results Five papers evaluating measurement properties of five self-reported health-related work-functioning instruments in CMD populations were included. There is little evidence available for the measurement properties of the identified instruments in this population, mainly due to low methodological quality of the included studies.

Conclusions The available evidence on measurement properties is based on studies of poor-to-fair methodological quality. Information on a number of measurement properties, such as measurement error, content validity, and cross-cultural validity is still lacking. Therefore, no evidence-based decisions and recommendations can be made for the use of health-related work functioning instruments. Studies of high methodological quality are needed to properly assess the existing instruments' measurement properties.

Key terms mental health; presenteeism; psychometrics; validation.

During the past decade, common mental disorders (CMD), such as depressive, anxiety, and adjustment disorders, have emerged as a major public and occupational health problem in many countries. On the societal level, CMD contribute to productivity loss, sickness absence, early retirement, and work disability (1–6). On the individual level, CMD cause not only suffering, but also have a negative impact on social relationships and social and work functioning (6). Several studies have shown a relationship between CMD and work performance (7, 8), and it has been estimated that the costs of lost productivity at work for CMD are much higher than those for absenteeism (7, 9–11).

In the field of occupational health practice and research, instruments are needed to assess lost productivity at work, monitor abilities to accomplish the work role, and evaluate interventions designed to improve work functioning (12, 13). Several self-reported questionnaires have been developed to measure the influence of health on functioning at work [for reviews see for example (12, 14–19)]. The joint influence of work and health determines an individual's work functioning. Two aspects of

1 University Medical Center Groningen, University of Groningen, Groningen, the Netherlands.
2 Department of Epidemiology & Biostatistics and EMGO Institute for Health & Care Research, VU University MC, Amsterdam, the Netherlands.
3 Institute for Work & Health, Toronto, ON, Canada.
4 University of Texas School of Public Health, Health Science Center at Houston, Houston, TX, USA.

Correspondence to: FI Abma, Department of Health Sciences, Community and Occupational Medicine, University Medical Center Groningen, University of Groningen, Antonius Deusinglaan 1, FA10, Room 610, 9713 AV Groningen, the Netherlands. [E-mail: f.i.abma@umcg.nl]

work functioning can be described. The first category deals with the economic consequences of health conditions, such as self-reported loss of productivity on the job. The second category deals with the reported limitations to meet work demands (13). Recently, a review by Nieuwenhuijsen et al (20) provided a narrative overview of work functioning in CMD populations, including instruments, dimensions of work functioning, and measurement properties. In this review, a systematic assessment of the methodological quality of the validation studies was not performed. However, to conduct an evidence synthesis, a systematic quality assessment is crucial because the results of poor quality studies may be biased (21).

Practitioners and researchers should make evidence-based decisions on which instrument to use. To select appropriate instruments for use in occupational health practice and research, the measurement properties (eg, reliability, validity, responsiveness) must be evaluated. If, for example, these instruments are used to evaluate interventions, it is important to know whether the instrument is able to detect changes over time. The synthesized evidence provided in systematic reviews on measurement properties should be used for the selection of instruments. A recent review of the measurement properties of health-related work-functioning instruments in populations with musculoskeletal disorders included a quality assessment, but a validated quality assessment tool was not used (19).

This review focuses on the measurement properties of self-reported health-related work-functioning instruments in CMD populations. Most of these instruments are designed for a broader population, but many are also used in CMD populations. However, the evidence for this use remains unclear. Therefore, the objective of this study was to appraise critically and compare the measurement properties of the identified self-reported health-related work functioning instruments in CMD populations.

Methods

Search strategy

The following electronic databases were searched: Embase, PsycInfo (EBSCOhost), and MEDLINE (PubMed). The search strategy consisted of search terms for the following characteristics, combined with "AND": (i) construct of interest (health-related work functioning); (ii) target population (CMD); and (iii) studies on measurement properties. Some examples of search terms that were used include: work performance, work functioning, work limitations, mental disorders, anxiety disorders, depressive disorder, and adjustment disorder. The complete search strategy can be found in the Appendix (http://www.sjweh.fi/data_repository.php). To identify studies on measurement properties in PubMed, we used a sensitive filter specially designed for identifying studies on measurement properties of patient-reported outcomes (22). This filter was adapted for searches in PsycInfo and Embase. No restrictions were made on the year of publication or language. Names of the retrieved instruments were used for further searches in the databases. Reference lists were screened to identify additional relevant studies.

Selection criteria

Health-related work-functioning instruments measure the influence of health on functioning at work. These types of instruments ask the respondent to rate the influence of his/her health status on his/her work functioning. Health-related work functioning is the ability of a worker to accomplish work demands given his or her state of health. In this review, we included instruments that both evaluate health-related work functioning and are from the worker's perspective (ie, self-reported). Instruments based on a single item, those measuring absenteeism only, or those whose work definitions included house and school work in addition to (paid) work were excluded.

Papers were included that: (i) mainly focused on the development and/or evaluation of the measurement properties of a self-reported health-related work-functioning instrument; (ii) were conducted in a population with CMD [including: depressive, anxiety, and adjustment disorders; diagnoses based on validated questionnaires, diagnostic interviews, or Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria]; (iii) were fulltext original papers (case studies, abstracts, letters to the editor, book chapters, conference proceedings, and unpublished papers were excluded). More severe psychiatric disorders, such as bipolar depression, psychosis, and schizophrenia were excluded.

Two independent reviewers screened titles and abstracts using the inclusion criteria. If there was any doubt as to whether the paper met the criteria, consensus was reached among the reviewers. Two independent reviewers reviewed the fulltext papers for inclusion. If necessary, a third independent reviewer was consulted.

Measurement properties

For the critical appraisal of the measurement properties, the consensus-based standards for the selection of health status measurement instruments (COSMIN) taxonomy was used. The COSMIN taxonomy was developed to provide an overview of the relevant measurement properties for health-related patient-reported outcomes and is based on international consensus (21, 23). According to the taxonomy, the measurement properties cover three quality domains: reliability, validity, and responsiveness (23). In addition, the interpretability of results is described.

Reliability is the extent to which scores for individuals who have not changed are the same for repeated measurement under several conditions [eg, using different sets of items from the same questionnaire (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons on different occasions (intra-rater)] (23). The reliability domain contains the following measurement properties: (i) internal consistency: the degree of interrelatedness among the items (expressed by Cronbach's α or Kuder-Richardson Formula (KR-20) (21, 23); when internal consistency is relevant, factor analysis or principal component analysis should be applied to determine whether the items form one or more than one scale (24); (ii) reliability: the proportion of the total variance in the measurements that reflects the "true" differences among individuals, including test-retest, inter- and intra-observer reliability [this aspect is reflected by the intraclass correlation coefficient (ICC) or Cohen's κ] (23, 25); (iii) measurement error: the systematic and random error of an individual's score that is not attributed to true changes in the construct to be measured, expressed by the standard error of measurement (SEM). The SEM can be converted into the smallest detectable change (SDC) (26). Changes exceeding the SDC can be labelled as change beyond measurement error. Another approach is to calculate the limits of agreement (LoA) (27). For determining the adequacy of measurement error, the SDC and/or LoA is related to the minimal important change (MIC) (28).

Validity is described as the degree to which an instrument measures the construct(s) it purports to measure (23). The validity domain contains three measurement properties: (i) content validity: the degree to which the content of the instrument is an adequate representative of the construct to be measured (including face validity). Content validity is an assessment of whether all items are relevant for the construct, aim and target population, and if no important items are missing (preferably by the target group) (29); (ii) construct validity is divided into three aspects: (a) structural validity: the degree the instrument scores are an adequate reflection of the construct's dimensionality. Factor analysis should be performed to confirm the number of subscales present; (b) hypotheses testing: the degree to which the instrument scores are consistent with hypotheses based on the assumption that the instrument validly measures the construct. Many different hypotheses can be formulated and tested (eg, the extent scores on a particular instrument relate to scores on other instruments or expected differences in scores between "known" groups. It is important in hypotheses testing to state hypotheses a priori, clearly indicating both direction and magnitude of the correlation or difference (29). For example, higher correlations are expected with similar constructs and variables, and lower correlations with dissimilar constructs and variables; (c) cross-cultural validity: the degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the original version of the instrument; (iii) criterion validity: the degree to which the scores of an instrument are an adequate reflection of a "gold standard". Since no real gold standard is available for measuring health-related work functioning, we will not evaluate criterion validity (29).

Responsiveness is described as the ability of an instrument to detect change over time in the construct to be measured (23). The responsiveness domain is considered an aspect of validity in a longitudinal context (29). Therefore, appropriate measures to evaluate responsiveness are the same as those for hypotheses testing and criterion validity. The only difference here is that hypotheses should focus on the change score of an instrument. Another approach is to determine the area under the receiver operator characteristic curve (AUC).

Interpretability is the degree to which one can assign qualitative meaning – that is clinical or commonly understood connotations – to an instrument's quantitative scores or change in scores (23). Investigators should provide information about clinically meaningful differences in scores between subgroups, floor and ceiling effects, and MIC. Although interpretability is not a measurement property, it is considered to be an important characteristic of an instrument.

Data extraction and description of the instruments

Two independent reviewers performed the data extraction. The retrieved instruments are described based on the information in original publications and the papers included in the review. The content, domains, target population, number of items, response options, and time to administer are presented (23). The measurement properties are presented as studied in the included papers.

Quality assessment

Assessing the quality of the included studies (on the measurement properties of the instruments) is an essential step of a systematic review of measurement properties. If the quality of a study is appropriate, the results are valid and the measurement instrument can be a useful tool in clinical practice or research. However, when the quality of a study is inadequate, the results cannot be trusted and the quality of the measurement instrument under study remains unclear. The methodological quality assessment was conducted using the COSMIN checklist (21). The COSMIN checklist is used to rate the quality of studies on one or more of the nine measurement properties (internal consistency, reliability, measurement error, content validity, structural validity, hypothesis testing, cross-cultural validity, criterion validity, and responsiveness) and the

quality of studies on interpretability. For each study on a measurement property, the methodological quality for that particular measurement property is rated by a series of items on a 4-point rating scale (poor, fair, good, excellent), which is an additional feature of the COSMIN checklist (30). These items rate for example the used sample sizes, the description of used comparator measures, how missing items were handled, and whether the used methods and statistics were appropriate. Per measurement property, an overall score for the methodological quality of a paper is determined by taking the lowest rating of any of the items per measurement property (poor to excellent). For example, if no description of the comparator instruments is provided for hypotheses testing, this item is rated "poor". Even though all others items for hypotheses testing may be rated "excellent", the methodological quality for hypotheses testing is rated "poor".

To rate the results of the measurement properties as positive, negative, or indeterminate, criteria were used based on Terwee et al (24). The criteria are presented in appendix 2. For example, for internal consistency, a positive (+) rating is given if Cronbach's α is ≥ 0.70, a negative (-) rating is given for <0.7, and an indeterminate (?) rating is given if no Cronbach's α is presented.

Two independent reviewers performed an assessment of methodological quality per paper. When two reviewers disagreed, there was a discussion to reach consensus. If necessary, a third reviewer made the decision.

Best evidence synthesis

A best evidence synthesis for each instrument was performed to summarize the total body of evidence for each measurement property, taking into account the number of studies, their quality, and the consistency of their results. Therefore, for each instrument, the rating of the methodological quality is combined with the rating of the measurement properties. The following criteria were used: a strong level of evidence = consistent findings in multiple studies of good methodological quality *or* in one study of excellent methodological quality; a moderate level of evidence = consistent findings in multiple studies of fair methodological quality *or* in one study of good methodological quality; a limited level of evidence = one study of fair methodological quality; and conflicting level of evidence = conflicting findings. When there were only studies of poor methodological quality, an unknown level of evidence was noted.

Results

Figure 1 shows the results of the selection procedure. The search resulted in 1630 references, after removing duplicates. Names of the retrieved instruments were used for further searches in the electronic databases. The most common reasons for exclusion at this stage were either that the paper was not about a health-related work functioning instrument that fit the inclusion criteria or it was not a validation study. The search strategy used to identify validation studies was very sensitive and resulted in a large number of references, including studies that were not validation studies.

Based on title and abstract, 71 fulltext papers were selected. The most common reasons for exclusion based on the fulltext were that the paper did not state a main aim to validate a self-reported health-related work-functioning instrument that fit the inclusion criteria or the study did not consist of a population with CMD. Finally, five papers evaluating five different self-reported health-related work-functioning instruments were included (31–35). Refer-

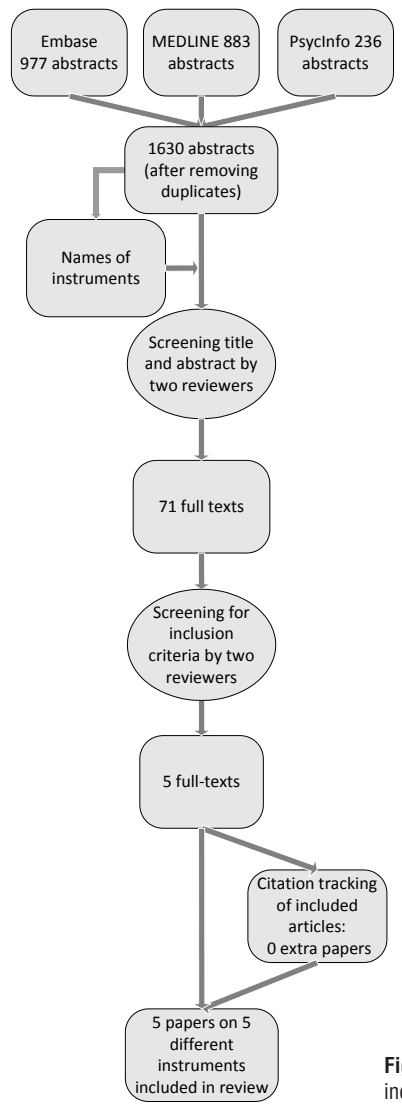

Figure 1. Flowchart of inclusion

ences of the retrieved papers were screened for additional relevant studies. No additional publications were found.

Identified instruments

Five different self-reported health-related work-functioning instruments are included: (i) the Endicott Work Productivity Scale (EWPS) (31), (ii) the Work Limitations Questionnaire (WLQ) (36, 37), (iii) the Stanford Presenteeism Scale (SPS) (38), (iv) the Work Performance Scale of the Functional Status Questionnaire (WPS) (39, 33), and (v) the Lam Employment Absence and Productivity Scale (LEAPS) (35).

The aim of all the instruments is to measure the degree to which health problems affect an individuals' work functioning. The EWPS and WLQ were developed for populations with a wide variety of both health conditions and jobs. The WPS and SPS were developed for working populations and the LEAPS was specially designed for a depressed (working) population. The WPS, SPS, and LEAPS are short questionnaires (between 6 and 7 items on work functioning) compared to the EWPS and WLQ (both 25 items). Table 1 presents an overview and description of the identified instruments.

Identified papers

Endicott et al (31) presented EWPS and investigated the internal consistency, test-retest reliability, hypotheses testing, and responsiveness in a population with major depression (diagnoses based on DSM-III-R criteria) and community subjects. The patients were recruited from an outpatient facility of a psychiatric institute. Uguz et al (32) translated the EWPS to Turkish and evaluated the internal consistency, test-retest reliability, and hypotheses testing in a population of depressed patients [diagnoses based on Hamilton Depression Rating Scale (HAM-D) and Structured Clinical Interview for DSM-IV (SCID)] and a community sample. Erickson et al (33) evaluated the EWPS, WLQ, and WPS in a population with anxiety disorders (diagnosis based on a clinical interview and consensus review by interdisciplinary team). They examined internal consistency, hypotheses testing, and responsiveness. Sanderson et al (34) evaluated the internal consistency, hypotheses testing, and responsiveness of the WLQ and SPS-6 in a population of call-center workers. They used the Patient Health Questionnaire to identify workers with depression and anxiety. For LEAPS, Lam et al (35) investigated the internal consistency, structural validity, and hypotheses testing in a population with major depressive disorder (diagnosis based on DSM-IV, clinical interview, symptom checklist and available medical records). All participants in the included papers were working. Table 2 shows an overview of the identified study populations.

Table 3 shows the measurement properties per instrument as reported in the included papers. Table 4 presents the methodological quality of each paper per measurement property and instrument, as rated with the COSMIN checklist. In table 5 the combined result of the methodological quality of the papers with the rating of the measurement properties are presented as a best evidence synthesis per measurement property of each instrument.

Endicott Work Productivity Scale (EWPS)

Reliability. Internal consistency was studied in the three papers that evaluated the measurement properties of the EWPS (31–33). Although the Cronbach's α were high (table 3), the studies were of poor methodological quality due to small samples and the fact that no factor analyses were performed in this population. Endicott et al (31) and Uguz et al (32) evaluated the test-retest reliability. Endicott et al used a small sample size (N=16), thus the paper was of poor methodological quality. The Uguz et al paper was of fair methodological quality due to moderate samples, and it was unclear if patients were stable. Measurement error was not studied in any of the papers.

Validity. Hypotheses testing was performed in all papers. Although no clear hypotheses were stated a priori, it was possible to deduce what was expected. As is shown in table 3, the EWPS was correlated with several other measures [eg, clinical state (32), clinical state at intake and endpoint (31), Symptom Checklist 90 (SCL-90) (31), SF-36 emotional and physical roles, the work-item of the Sheehan Disability Scale (SDS) (33), and SF-36 social functioning subscale (32)]. Endicott (31) did not properly describe the constructs and instruments used and, therefore, the paper was of poor methodological quality. The others two papers (32, 33) were of fair methodological quality. The cross-cultural validity was not assessed, although the translation process was described by Uguz et al (32). Content validity and structural validity were not studied.

Responsiveness. Responsiveness over time was evaluated by Endicott (correlations with change scores in HAM-D) and Erickson (effect size calculated between change scores in two groups based on change in severity of illness) as shown in table 3. Papers were of poor methodological quality because of small sample size (33) and a lack of important information on time interval and comparator instruments (31).

Interpretability. Regarding the interpretability, floor or ceiling effects, and MIC were not studied. Scores and change scores were presented for relevant subgroups (31–33).

Best evidence synthesis. The evidence synthesis for the EWPS (table 5) resulted in unknown evidence (?) for

Table 1. Description of identified instruments.

Instrument and reference(s)	Content/aim and target population	Domains, number of items, and example item	Response options and recall period	Scoring and time to administer
Endicott Work Productivity Scale (EWPS). Endicott et al, 1997 (31)	Designed to assess the degree to which a medical condition, such as a depressive disorder, affects the work functioning of a subject. The EWPS is designed to assess subjects with a wide variety of mental and medical disorders working in a wide variety of job settings including self-employment.	1 domain: work productivity 25 items plus additional items on expected working hours, hours worked, and reason for working less (if applicable). Example item: "During the past week, how frequently did you arrive at work late or leave early"	5-point scale rating how often the behavior or feeling or attitude has been manifested during the past week: 0=never,1=rarely, 2=sometimes, 3=often, 4=almost always Recall period: past week	Total score ranges from 0 (best possible score) to 100 (worst possible score). "A brief self-report questionnaire"
Work Limitations Questionnaire (WLQ). Lerner et al, 2001 (36, 37)	To measure the degree to which chronic health problems interfere with the ability to perform job roles (on a demand-level). Measuring the on-the-job impact if chronic health problems and/or treatment ("work limitations"). It can be used to identify both the magnitude and type of impact that health problems are having in the workplace. (Employed individuals with) chronic health problems in several different jobs/work conditions	4 domains: time scheduling demands (5 items); physical demands (6 items); mental-interpersonal demands (9 items); output demands (5 items): total 25 items. Example item: "In the past two weeks, how much of the time did your physical health or emotional problems make it difficult for you to stick to a routine or schedule?"	5-point rating scale: 0=all of the time (100%) 1=a great deal of the time 2=some of the time (~50%) 3=a slight bit of the time 4=none of the time (0%) Extra option: "not applicable to my job" Recall period: past 2 weeks is recommended, but 4 weeks also allowed	Total scores are computed as the mean of the non-missing responses and are converted to 0 (not limited) to 100 (limited all of the time) ("not applicable to my job" is scored as missing). "Easy self-report"
Stanford Presenteeism Scale, 6 items (SPS-6) Koopman et al, 2002 (38)	A presenteeism scale evaluating the impact of health problems on individual performance and productivity. Working populations	1 domain: one total score with two factors (completing work and avoiding distraction): 6 items. Example item: "Despite having my (health problem), I was able to finish hard tasks in my work"	5-point scale to agree/disagree with statement, with: 1=strongly disagree 2=somewhat disagree 3=uncertain 4=somewhat agree 5=strongly agree Recall period: past month	Total score is the sum of the values. A high score indicated a high level of presenteeism [ie, a greater ability to concentrate on and accomplish work despite health problem(s)]. No information available on time to administer
Work Performance Scale (WPS) Erickson et al, 2009 (33) Jette et al, 1986 (39)	From the Functional Status Questionnaire, the WPS aims to assess changes in the job-related to health, ability to perform tasks, time required to perform tasks and interpersonal relationships. Employed during previous month	1 domain: work performance (6 items) Example item: "If you were employed the last month, how was your work performance done as much work as others in similar jobs?"	Original scale is 4-point rating scale (Erickson uses a 5-point rating scale, with low scores reflecting impairment): 1=all the time; 2=most of the time; 3=some of the time; 4=none of the time. Recall period: past month.	The scores are transformed to a single scale score from 0–100, with 100 indicating maximum functional ability. No information available on time to administer
Lam Employment Absence and Productivity Scale (LEAPS). Lam et al, 2009 (35)	Designed to assess work functioning and impairment in a clinically depressed population. A depressed population (major depressive disorder) (working)	2 domains: productivity (3 items) and troublesome symptoms (4 items); 7 items plus 3 items on occupation, hours of work, and hours missed from work. Example item: "Over the past two weeks, how often were you bothered by getting less work done"	5-point scale: 0=None of the time (0%) 1=some of the time (25%) 2=half of the time (50%) 3=most of the time (75%) 4=all of the time (100%) Recall period: past 2 weeks	Score 0–4 respectively. Total score range from 0–28. No information available on time to administer

internal consistency; unknown evidence (?) for reliability (test-retest); limited positive evidence (+) for hypotheses testing (two studies of fair methodological quality and 75% of the results are in accordance with hypotheses); and unknown evidence (?) for responsiveness (two studies of poor methodological quality and one with limited positive evidence).

Work Limitations Questionnaire (WLQ)

Reliability. Internal consistency was studied by Sanderson et al (33) and Erickson et al (34). Although Cronbach's α were high in both studies (table 3), Sanderson et al reported no information on performed factor analysis and therefore the paper was of poor methodological quality. The Erickson et al paper was of fair methodological quality because the authors only refer to a study that performed factor analyses. Test-retest and measurement error were not studied in either paper.

Validity. Both studies performed hypotheses testing. Although Erickson did not state clear a-priori hypoth-

Table 2. Description of identified study populations. [DSM=Diagnostic and Statistical Manual of Mental Disorders; EWPS=Endicott Work Productivity Scale; HAM-D=Hamilton Depression Rating Scale; ICC=interclass correlation coefficient; LEAPS=Lam Employment Absence and Productivity Scale; SCID=Structured Clinical Interview for DSM-IV; SPS=Stanford Presenteeism Scale; WLQ=Work Limitations Questionnaire; WPS=Work Performance Scale.]

Paper	Instrument	Setting	Employment status	Country & language	Number of subjects (% female) & study population	Mean age (years)	SD
Endicott et al, 1997 (31)	EWPS	Outpatient facility of the New York State Psychiatric Institute	Currently working	US, language of questionnaire not stated	N=42 (50.0%). Major depression (DSM-III-R). A group within the sample had alcoholism as a comorbid mental disorder.	41	9.6
					N=66 (70.0%). Community subjects	.	.
Uguz et al, 2004 (32)	EWPS	Patients visiting psychiatric department of university hospital	Currently working	Turkey, Turkish	N=74 (70.3%) Depressed patients (HAM-D/ SCID interview DSM-IV)	.	.
					N=107 (60.7%) Community sample	.	.
Erickson et al, 2009 (33)	WLQ EWPS WPS	Patients seeking evaluation for anxiety treatment at the university of Michigan Anxiety Disorders Program	Work for pay >20 hours/week	US, English	N=41 (48.8%) Minimal to mild anxiety disordera	37.5	12.2
					N=40 (75.0%) Moderate to severe anxiety disorder [a]	34.2	9.8
Sanderson et al, 2007 (34)	WLQ SPS-6	Call center workers	Employment contract	Australia, language of questionnaire not stated	N=436 (77.1%) Community sample	.	.
					N=69, selected from community sample (NR) Depression and anxiety within community sample [b]	.	.
Lam et al, 2009 (35)	LEAPS	Patients attending a mood disorders clinic at a university teaching hospital	Paid work [(self)employed, either part-time or full-time]. Workers on short- or long-term disability are excluded	Canada, language of questionnaire not stated	N=234 (NR). Major depressive disorder (DSM-IV, clinical interview, symptom checklist and available medical records)	39.2	11.7

[a] Anxiety disorder (generalized anxiety, panic, obsessive-compulsive disorder, or social phobia). Diagnosis based on a 2-hour clinical interview and consensus review by inter-disciplinary team. Beck Anxiety Inventory was used for defining two severity groups.

[b] Diagnosis by depression and anxiety modules of the Patient Health Questionnaire

eses, it was possible to deduce what was expected. As is shown in table 3, the WLQ was correlated to several other measures [eg, SF-36 emotional and physical roles, the work-item of the SDS (33)] and comparisons between severity groups were made (33,34). Sanderson used small patient groups for the comparison and therefore the paper was of poor methodological quality. Erickson reported little information on the expectations and comparator instruments and therefore the paper was of fair methodological quality. The content validity, structural validity, and cross-cultural validity were not studied.

Responsiveness. To evaluate responsiveness, Sanderson et al compared the WLQ scores by symptom status at baseline, 6 months, and change scores between four groups. Erickson et al calculated effect sizes between change scores in two groups based on change in severity of illness (table 3).

Interpretability. Neither floor nor ceiling effects nor MIC were studied; however scores and change scores were presented for relevant subgroups (33, 34).

Best evidence synthesis. Evidence synthesis of the WLQ (table 5) resulted in limited positive evidence (+) for internal consistency (two studies with poor and fair methodological quality and Cronbach's α >0.80); limited positive evidence (+) for hypotheses testing (two studies with poor and fair methodological quality and 75% of the results were in accordance with hypotheses); and unknown evidence (?) for responsiveness based on two studies with poor methodological quality.

Stanford Presenteeism Scale 6-item scale (SPS-6)

Reliability. As shown in table 3, Sanderson et al (34) investigated the internal consistency but did not report on factor analysis in any population on the SPS-6. Therefore, the paper was of poor methodological quality. The test-retest and measurement error were not studied.

Validity. Hypotheses testing was performed by comparing different severity of depression groups at baseline (table 3). The content validity, structural validity, and cross-cultural validity were not studied.

Table 3a. Reported measurement properties per instrument: internal consistency, reliability and structural validity [a]. [EWPS=Endicott Work Productivity Scale; HAM-D=Hamilton Depression Rating Scale ; ICC=interclass correlation coefficient; LEAPS=Lam Employment Absence and Productivity Scale; MID=mental-interpersonal demands; OD=output demands; PCA=principal component analysis; PD=physical demands; SPS=Stanford Presenteeism Scale; TSD=time scheduling demands; WLQ=Work Limitations Questionnaire; WPS=Work Performance Scale]

Instrument and paper	Internal consistency		Reliability		Structural validity	
	Study population	Results	Study population	Results	Study population	Results
EWPS						
Endicott et al, 1997 (31)	N=108, total sample	α=0.93	N=16, subset of total sample	Test–retest ICC=0.92		
	N=42, patient group	α=0.93				
	N=66, community sample	α=0.92				
Uguz et al, 2004 (32)	N=74, patient group	α=0.90	N=30, subset of total sample	Test–retest: r=0.76		
	N=107, community sample	α=0.92				
Erickson et al, 2009 (33)	Total sample	α=0.95				
WLQ-25						
Erickson et al, 2009 (33)	Total sample	TSD α=0.91 PD α=0.92 MID α=0.92 OD α=0.93				
Sanderson et al, 2007 (34)	Total sample	α>0.89				
SPS-6						
Sanderson et al 2007 (34)	Total sample	α=0.70				
WPS						
Erickson et al, 2009 (33)	Total sample	α=0.82				
LEAPS						
Lam et al, 2009 (35)	Total sample	Total scale α=0.89			N=234, total sample	PCA with Varimax rotation: two expected factors found, together 75% explained variance.

[a] Criterion validity was not evaluated. No evidence available for measurement error, content validity and interpretability.

Table 3b. Reported measurement properties per instrument: hypotheses testing and responsiveness [a]. [DSM=Diagnostic and Statistical Manual of Mental Disorders; EWPS=Endicott Work Productivity Scale; HAM-D= Hamilton Depression Rating Scale; HPQ= World Health Organization Health and Work Performance Questionnaire; LEAPS=Lam Employment Absence and Productivity Scale; MID=mental-interpersonal demands; OD=output demands; PCA=principal component analysis; PD=physical demands; SCL=symptom checklist; SD= standard deviation; SPS=Stanford Presenteeism Scale; TSD=time scheduling demands; WLQ=Work Limitations Questionnaire; WPS=Work Performance Scale]

Instrument and paper	Hypotheses testing		Responsiveness	
	Study population	Results	Study population	Results
EWPS [b]				
Endicott et al, 1997 (31)	N=42, major depression group	Intake: Correlations with HAM-D (r=0.27) and Global Clinical Index of severity (r=0.42) Endpoint: Correlations with HAM-D (r=0.61), Global Clinical Index of severity(r=0.46), and SCL-90 (r=0.50)	Patient group	Correlations change score with HAM-D (r=0.29)
	N=66, community sample	Intake: Correlations with Zimmerman total score (r=0.57) and SCL-90 (r=0.55)		
	Total sample	Patients had higher EWPS scores than community sample at both intake and endpoint.		
Uguzet al, 2004 (32)	N=74, depression group	Comparison of scores with Hamilton depression scale (r=0.52), SF-36 social functioning subscale (r=-0.43), clinical global impression severity scale (r=0.64)		
	N=181, total sample (depression group and control group)	Significant difference between patient group and control group (mean difference 24 points)		

(cont)

Table 3b. Continued

Erickson et al, 2009 (33)	N=76, anxiety disorder group	Correlation (r) with SF-36 emotional role: -0.63, SF-36 physical role: -0.23 and SDS-work item 0.63	N=38 Perceived improved severity of illness N=12 Perceived no change or worsening severity of illness Two change in severity of illness groups based on global improvement scale (CGI-I)	Group comparisons of change scores over 12 week: mean change/SD, P-value: No significant differences. Effect size 0.71.
	N=41 minimal-to-mild anxiety; N=40 moderate-to-severe anxiety. Two severity of illness groupsbased on Beck Anxiety Inventory	Group comparisons on mean/SD, P-value and effect size: higher scores for severe anxiety group, effect size -0.45.		
WLQ-25 [b]				
Erickson et al, 2009 (33)	N=76, anxiety disorder group	Correlation (r) of subscales with: (i) SF-36 physical role: TSD=-0.24, OD=-0.26, PD=-0.19, MID=-0.24; (ii) SF-36 emotional role: TSD=-0.69, OD=-0.65, PD=-0.23, MID=-0.74; (iii) SDS-work item: TSD=0.56, OD=0.62, PD=0.16, MID=0.65	N=38 Perceived improved severity of illness N=12 Perceived no change or worsening severity of illness Two change in severity of illness groups based on Global Improvement Scale (CGI-I)	Per subscale group comparisons of change scores over 12 week: mean change/SD, P-value: mixed results. Effect sizes: TSD=-0.35 OD=-0.86 PD=-0.01 MID=-1.03
	N=41 and N=40. Two severity of illness groups based on Beck Anxiety Inventory	Per subscale group comparison on mean/SD, P-value and effect size: higher limitations for severe anxiety. Effect sizes: TSD=-0.35, OD=-0.86, PD=-0.01, MID=-1.03		
Sanderson et al 2007 (34)	N=363: No depressive syndrome N=69: Any depressive syndrome N=24: Minor depressive syndrome N=25: Major depressive syndrome	Per subscale group comparisons: no depressive syndrome, minor depressive syndrome, and major depressive syndrome on mean and SD. TSD: (i) any versus no syndrome; B=17.4, SE=2.6, P<0.0001. Group without depression/anxiety had lower mean than group with any (less limitations); (ii) minor versus no syndrome; B=12.3, SE= 3.8 P=0.010. Group without depression/anxiety had lower mean than group with minor; (iii) major versus no syndrome; B=13.6, SE=4.8, P=0.019. Group without depression/anxiety had lower mean than group with minor OD: (i) any versus no syndrome; B=18.0, SE=2.8, P<0.0001. Group without depression/anxiety had lower mean than group with any; (ii) minor versus no syndrome; B=12.0, SE=3.1, P=0.004. Group without depression/anxiety had lower mean than group with minor; (iii) major versus no syndrome; B=16.6, SE=3.8, P=0.002. Group without depression/anxiety had lower mean than group with minor PD: (i) Any versus no syndrome; B=12.5, SE=1.2, P<0.0001. Group without depression/anxiety had lower mean than group with any; (ii) minor versus no syndrome; B=14.5, SE=3.3 P=0.002. Group without depression/anxiety had lower mean than group with minor; (iii) major versus no syndrome; B=-4.2, SE=3.0, P=0.20. Group without depression/anxiety had lower mean than group with minor MID: (i) any versus no syndrome; B=17.7, SE=1.9, P<0.0001. Group without depression/anxiety had lower mean than group with any; (ii) minor versus no syndrome; B=10.5, SE=2.4, P=0.002. Group without depression/anxiety had lower mean than group with minor; (iii) major versus no syndrome; B=17.5, SE=3.6, P=0.0009. Group without depression/anxiety had lower mean than group with minor.	N=174: Remained symptom free N=21: Onset of syndrome at 6 months N=20 Syndrome remitted at 6 months N=16 Syndrome persisted at six months	Per subscale group comparisons in presenteeism scores (mean/sd) by depression/anxiety syndrome status at baseline, 6 months and change score: Mixed results, nevertheless most are in the expected direction.
	N=427: total sample (call centre workers)	A linear regression model is applied: Relationships of subscales with specific DSM-IV depression symptoms at baseline evaluated in a regression model: mixed results		

(cont)

Table 3b. Continued

SPS-6 [b]				
Sanderson et al 2007 (34)	N=363: No depressive syndrome N=69: Any depressive syndrome N=24: Minor depressive syndrome N=25: Major depressive syndrome	Group comparisons: no depressive syndrome, minor depressive syndrome, and major depressive syndrome are on mean and SD. (i) any versus no syndrome; B=-3.8, SE=0.4, P<0.0001. Group without depression/anxiety had higher mean than group with any (better functioning; (ii) minor versus no syndrome; B=-3.1, SE=0.8, P=0.004. Group without depression/anxiety had higher mean than group with minor; (iii) major versus no syndrome; B=-1.1, SE=0.9, P=0.25. Group without depression/anxiety had higher mean than group with minor.	N=174: Remained symptom free N=21: Onset of syndrome at 6 monthts N=20 Syndrome remitted at 6 months N=16 Syndrome persisted at six months	Group comparisons in presenteeism scores (mean/SD) by depression/anxiety syndrome status at baseline, 6 months and change score. Almost all results are in the expected direction, although not all large differences/significant differences.
WPS [b]				
Erickson et al, 2009 (33)	N=76, anxiety disorder group	Correlation (r) with: (i) SF-36 emotional role: 0.66; (ii) SF-36 physical role: 0.34; (iii) SDS-work item: -0.69	N=38 Perceived improved severity of illness N=12 Perceived no change or worsening severity of illness Two change in severity of illness groups based on Global Improvement Scale (CGI-I)	Group comparisons of change scores over 12 week: mean change/SD, P-value: no significant results. Effect size 0.49.
	N=41 and N=40. Two severity of illness groups based on Beck Anxiety Inventory	Group comparisons: mean/SD, P-value: lower scores for severe anxiety. Effect size=-0.45 (worse functioning)		
LEAPS [b]				
Lam et al, 2009 (35)	N=234, major depression group	Correlation (r) of total score and productivity subscale with: (i) SDS-work item (0.63, 0.50); (ii) HPQ global work performance (-0.79, -0.85); (iii) HPQ productivity score (-0.70, -0.77); (iv) percent of missed hours in past two weeks (0.41, 0.45). Comparison between severity groups and their scores on the total score and productivity subscale. The severe categories showed higher scores (one way ANOVA) (worse functioning).		

[a] Criterion validity was not evaluated. No evidence available for measurement error, content validity and interpretability.

Responsiveness. To evaluate the responsiveness, the authors compared the SPS-6 scores by symptom status at baseline, 6 months, and change scores in four groups (table 3).

Interpretability. Differences in scores and change scores for relevant subgroups were presented; neither floor nor ceiling effects nor MIC were studied (34).

Best evidence synthesis. Although all results were in the expected directions for internal consistency, hypotheses testing, and responsiveness, there was unknown evidence (?) because of the paper's poor methodological quality (small groups in analyses) (table 5).

Work Performance Scale (WPS)

Reliability. As shown in table 3, Erickson et al (33) evaluated the internal consistency but did not report on factor analysis in any population on the WPS. Therefore, the paper was of poor methodological quality. The test-retest and measurement error were not studied.

Validity. Hypotheses testing was performed by correlating the WPS to several other measures (eg, SF-36 emotional and physical roles, the work-item of the SDS) and a comparison between two severity groups was made (table 3). No clear a priori hypotheses were stated and little information was reported on comparator instruments. Therefore, the paper was of fair methodological quality. The content validity, structural validity, and cross-cultural validity were not studied.

Responsiveness. The responsiveness was evaluated by an effect size between change scores in two groups based on change in severity of illness (table 3).

Interpretability. Differences in scores and change scores for relevant subgroups were presented; no floor or ceiling effects and MIC were studied (33).

Best evidence synthesis. Because of the poor methodological quality (small sample sizes), there was unknown evidence (?) for the internal consistency and responsiveness of the WPS (table 5). For hypotheses testing limited positive evidence (+) was found.

Table 4. Methodological quality of each paper per measurement property and instrument [a] [EWPS=Endicott Work Productivity Scale; LEAPS=Lam Employment Absence and Productivity Scale; SPS=Stanford Presenteeism Scale; WLQ=Work Limitations Questionnaire; WPS=Work Performance Scale.]

Instrument and paper	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypotheses testing	Cross-cultural validity	Responsiveness	Interpretability
EWPS									
Endicott et al, 1997 (31)	Poor	Poor	·	·	·	Poor	·	Poor	·
Uguz et al, 2004 (32)	Poor	Fair	·	·	·	Fair	·	·	·
Erickson et al, 2009 (33)	Poor	·	·	·	·	Fair	·	Poor	·
WLQ-25									
Erickson et al, 2009 (33)	Fair	·	·	·	·	Fair	·	Poor	·
Sanderson et al, 2007 (34)	Poor	·	·	·	·	Poor	·	Poor	·
SPS-6									
Sanderson et al, 2007 (34)	Poor	·	·	·	·	Poor	·	Poor	·
WPS									
Erickson et al, 2009 (33)	Poor	·	·	·	·	Fair	·	Poor	·
LEAPS									
Lam et al, 2009 (35)	Poor	·	·	·	Fair	Fair	·	·	·

[a] Criterion validity was not evaluated. The methodological quality was assessed using the COSMIN checklist.

Table 5. Quality of measurement properties per instrument based on a best evidence synthesis of the combined information from all studies a [EWPS=Endicott Work Productivity Scale; LEAPS=Lam Employment Absence and Productivity Scale; SPS=Stanford Presenteeism Scale; WLQ=Work Limitations Questionnaire; WPS=Work Performance Scale; ++ = moderate positive evidence; + limited positive evidence; ? = unknown, due to poor methodological quality]

Measures	Internal Consistency	Reliability	Measurement error	Content validity	Structural validity	Hypotheses testing	Cross-cultural validity	Responsiveness
EWPS	?	?	·	·	·	++	·	?
WLQ	+	·	·	·	·	+	·	?
SPS-6	?	·	·	·	·	?	·	?
WPS	?	·	·	·	·	+	·	?
LEAPS	?	·	·	·	+	+	·	·

[a] Criterion validity was not evaluated.

Lam Employment Absence & Productivity Scale (LEAPS)

Reliability. Lam et al (35) presented a new instrument, LEAPS, and investigated the internal consistency of the total scale (table 3). However, no Cronbach's α of the subscales were available, therefore the paper was of poor methodological quality.

Validity. Structural validity was studied by performing a factor analysis (principal component analysis with varimax rotation). The authors did not report how missing items were handled, resulting in fair methodological quality. Hypotheses testing was performed by correlating the LEAPS to several other measures (eg, the SDS work item, HPQ global work performance, HPQ productivity score, and % missed hours at work in past two weeks) and a comparison between five severity groups was made (table 3). Little information was provided on the a priori expectations; the paper used comparator instruments and was, therefore, of fair methodological quality.

Responsiveness. This domain was not studied.

Interpretability. Neither floor nor ceiling effects nor MIC were studied (35).

Best evidence synthesis. Due to the poor methodological quality, there was unknown evidence (?) for the internal consistency of the LEAPS (table 5). For structural validity and responsiveness, limited positive evidence (+) was found (fair methodological quality and positive results).

Discussion

This systematic review was conducted to identify the measurement properties of self-reported health-related work-functioning instruments among workers with common mental disorders, taking the methodological quality of the studies into account in a best evidence synthesis.

Five papers reporting on the measurement properties of five self-reported health-related work-functioning instruments were included. The results show that there is little evidence available for the measurement properties of the identified instruments in this population, mainly due to the poor-to-fair methodological quality of the included studies.

None of the five identified instruments showed satisfactory results for all measurement properties. The internal consistency of all instruments was evaluated (all Cronbach's α >0.70), as was construct validity by means of hypotheses testing: comparison of severity groups and correlations with other constructs. Test-retest reliability was only studied for the EWPS, in both the English (31) and Turkish (32) versions, with positive result. Although responsiveness was evaluated for four out of five instruments (EWPS, WLQ, SPS-6, and WPS), the results are difficult to interpret due to small sample sizes and inappropriate methods (31, 33, 34). Structural validity was evaluated for the LEAPS only (35). Measurement error, content validity, cross-cultural validity, and interpretability were either not studied or not adequately described in the included studies. Larger, well-designed validation studies in CMD populations are needed to provide more evidence for the measurement properties of health-related work-functioning instruments. In particular, large validation studies that include several of these instruments, in order to evaluate and compare the measurement properties, are needed.

Although the overall evidence for the measurement properties for all instruments is low, this does not imply that the instruments do not have good measurement properties. For example, the reported Cronbach's α of all instruments and subscales were >0.70, but for most instruments no factor analysis was performed in the study population to assess the unidimensionality. If there is no evidence that the scales are unidimensional, the Cronbach's α cannot be properly interpreted (40). Moreover, the focus of this review is on CMD populations, while the instruments also have been used and validated in other populations (12, 14–19). For example, the SPS-6, and the WLQ-16 were included among other instruments in a validation study performed among workers with shoulder or elbow disorders (18) and the WLQ-25, EWPS, and SPS-6 were included among other instruments in a validation study of a rheumatoid arthritis and osteoarthritis population (15). All instruments showed satisfactory measurement properties.

The overall methodological quality of the included studies was poor to fair. Several reasons were found for these low ratings: often very small sample sizes were used in the analyses, especially in subgroup analyses. When validating instruments by means of hypotheses testing, it is important to formulate clear (a priori) hypotheses, stating the direction and magnitude of expected correlations or mean differences. In the present review, only Sanderson (34) and Lam (35) formulated hypotheses in their papers. Another reason for poor methodological quality was the lack of information. For example, studies failed to report on how missing items were handled, time intervals for test-retest and responsiveness were not stated, and comparative instruments used in hypotheses testing were not described. The evidence synthesis was performed per measurement property for each instrument to categorize the total body of evidence. It has to be noted, that if there is no evidence available, no rating can be made. This is different from unknown evidence (?), which is based on studies of poor methodological quality.

All included papers focused on workers and all included health-related work-functioning instruments were designed for use in working populations, often addressing a wide range of health conditions. Most identified papers included study populations in a clinical setting, ie, most participants were recruited in healthcare settings. One paper recruited in a workplace setting (34). Different instruments and classifications were used to diagnose CMD. Caution is needed before generalizing the results to the day-to-day practice of occupational physicians, who may use these instruments to monitor work functioning of workers with CMD in a workplace setting.

An asset of this study is that it used a rather strict set of inclusion criteria for self-reported health-related work-functioning instruments. Papers were only included if they clearly stated that their aim was to validate a specific instrument. Moreover, instruments were only included if they were self-reported and evaluated work functioning or effectiveness on the job. Instruments based on a single item, those measuring absenteeism only, or those whose work definitions included house and school work were excluded. A recent review showed that there are more work functioning instruments used in CMD populations than the five included in this review (20). It might therefore be possible that, due to these strict set of inclusion criteria, instruments or papers were excluded that are also of interest for this population. However, because of our strict focus, this review provides a clear overview of the available evidence in this field and reveals gaps in knowledge.

The COSMIN taxonomy and checklist were chosen for the critical appraisal of the measurement properties (21, 23, 29, 30). The taxonomy was developed to provide an overview of the relevant measurement properties for health-related, patient-reported outcomes, and is based on international consensus (23). The COSMIN taxonomy might contribute to a better understanding and less ambiguity in the terminology and definitions used in validation studies. The COSMIN checklist provided a structured procedure for the evaluation of the methodological quality of studies on measurement proper-

ties. Although the COSMIN-checklist-based evaluation revealed that studies had poor-to-fair methodological quality, this does not imply that the instruments do not have good measurement properties.

The current systematic review had a narrow set of inclusion criteria with a structured procedure for the quality assessment. The results clearly indicate that there is a need for more and better validation studies in CMD populations for health-related work-functioning instruments. The COSMIN checklist may be used as a guide for designing methodologically sound validation studies.

Concluding remarks

This systematic review provides an overview of the available evidence on the measurement properties of health-related work-functioning instruments in CMD populations. Most evidence is limited, with the construct validity – by means of hypothesis testing – having the highest level of evidence for all instruments. Information on a number of measurement properties, such as measurement error, content validity, and cross-cultural validity is still lacking. Also information on interpretability of the instruments is mostly lacking. Therefore, no evidence-based decisions and/or recommendations can be made for the use of health-related work-functioning instruments in CMD populations. For now, in determining which instrument to employ, users will have to base their decisions on the content of the instrument, the purpose of use, and the target population, in addition to the little evidence available. Studies of high methodological quality are needed to properly assess the existing instruments' measurement properties. We recommend using the COSMIN checklist in the design of these studies.

Acknowledgements

The authors would like to thank Minne Yildrim for her help with translating and assessing Uguz et al (32).

References

1. Henderson M, Glozier N, Holland Elliott K. Long term sickness absence. BMJ. 2005;330:802–3. doi:10.1136/bmj.330.7495.802
2. Bultmann U, Rugulies R, Lund T, Christensen KB, Labriola M, Burr H. Depressive symptoms and the risk of long-term sickness absence: a prospective study among 4747 employees in Denmark. Soc Psychiatry Psychiatr Epidemiol. 2006;41:875–80.
3. Karpansalo M, Kauhanen J, Lakka TA, Manninen P, Kaplan GA, Salonen JT. Depression and early retirement: prospective population based study in middle aged men. J Epidemiol Community Health. 2005;59:70–4. doi:10.1136/jech.2003.010702
4. Goetzel RZ, Hawkins K, Ozminkowski RJ, Wang S. The health and productivity cost burden of the "top 10" physical and mental health conditions affecting six large U.S. employers in 1999. J Occup Environ Med. 2003;45:5–14. doi:10.1097/00043764-200301000-00007
5. Laitinen-Krispijn S, Bijl RV. Mental disorders and employee sickness absence: the NEMESIS study. Netherlands Mental Health Survey and Incidence Study. Soc Psychiatry Psychiatr Epidemiol. 2000;35:71–7. doi:10.1007/s001270050010
6. Hirschfeld RM, Montgomery SA, Keller MB, Kasper S, Schatzberg AF, Moller HJ, et al. Social functioning in depression: a review. J Clin Psychiatry. 2000;61:268–75. doi:10.4088/JCP.v61n0405
7. Lerner D, Henke RM. What does research tell us about depression, job performance, and work productivity? J Occup Environ Med. 2008;50:401–10. doi:10.1097/JOM.0b013e31816bae50
8. Sanderson K, Andrews G. Common mental disorders in the workforce: recent findings from descriptive and social epidemiology. Can J Psychiatry. 2006;51:63–75.
9. Loeppke R, Taitel M, Haufle V, Parry T, Kessler RC, Jinnett K. Health and productivity as a business strategy: a multiemployer study. J Occup Environ Med. 2009;51:411–28. doi:10.1097/JOM.0b013e3181a39180
10. Goetzel RZ, Long SR, Ozminkowski RJ, Hawkins K, Wang S, Lynch W. Health, absence, disability, and presenteeism cost estimates of certain physical and mental health conditions affecting U.S. employers. J Occup Environ Med. 2004;46:398–412. doi:10.1097/01.jom.0000121151.40413.bd
11. Collins JJ, Baase CM, Sharda CE, Ozminkowski RJ, Nicholson S, Billotti GM, et al. The assessment of chronic health conditions on work performance, absence, and total economic impact for employers. J Occup Environ Med. 2005;47:547–57. doi:10.1097/01.jom.0000166864.58664.29
12. Amick BC,III, Lerner D, Rogers WH, Rooney T, Katz JN. A review of health-related work outcome measures and their uses, and recommended measures. Spine. 2000;25:3152–60. doi:10.1097/00007632-200012150-00010
13. Amick BC,III, Gimeno D. Measuring Work Outcomes with a focus on Health-Related Work Productivity Loss. In: Wittink H, Carr D, editors. Pain Management: Evidence, Outcomes, and Quality of Life: A Sourcebook. Amsterdam: Elsevier; 2008. p. 329–43.
14. Loeppke R, Hymel PA, Lofland JH, Pizzi LT, Konicki DL, Anstadt GW, et al. Health-related workplace productivity measurement: general and migraine-specific recommendations from the ACOEM Expert Panel. J Occup Environ Med. 2003;45:349–59. doi:10.1097/01.jom.0000063619.37065.e2
15. Beaton DE, Tang K, Gignac MA, Lacaille D, Badley EM,

Anis AH, et al. Reliability, validity, and responsiveness of five at-work productivity measures in patients with rheumatoid arthritis or osteoarthritis. Arthritis Care Res. 2010;62:28–37. doi:10.1002/acr.20011

16. Lofland JH, Pizzi L, Frick KD. A review of health-related workplace productivity loss instruments. Pharmacoeconomics. 2004;22:165–84. doi:10.2165/00019053-200422030-00003

17. Prasad M, Wahlqvist P, Shikiar R, Shih YC. A review of self-report instruments measuring health-related work productivity: a patient-reported outcomes perspective. Pharmacoeconomics. 2004;22:225–44. doi:10.2165/00019053-200422040-00002

18. Tang K, Pitts S, Solway S, Beaton D. Comparison of the psychometric properties of four at-work disability measures in workers with shoulder or elbow disorders. J Occup Rehabil. 2009;19:142–54. doi:10.1007/s10926-009-9171-6

19. Williams RM, Schmuck G, Allwood S, Sanchez M, Shea R, Wark G. Psychometric Evaluation of Health-Related Work Outcome Measures For Musculoskeletal Disorders : A Systematic Review. J Occup Rehabil. 2007;17:504–21. doi:10.1007/s10926-007-9093-0

20. Nieuwenhuijsen K, Franche RL, van Dijk FJ. Work functioning measurement: tools for occupational mental health research. J Occup Environ Med. 2010;52:778–90. doi:10.1097/JOM.0b013e3181ec7cd3

21. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19:539–49. doi:10.1007/s11136-010-9606-8

22. Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res. 2009;18:1115–23. doi:10.1007/s11136-009-9528-5

23. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63:737–45. doi:10.1016/j.jclinepi.2010.02.006

24. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60:34–42. doi:10.1016/j.jclinepi.2006.03.012

25. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. New York: Oxford University Press; 2003xii, 283.

26. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59:1033–39. doi:10.1016/j.jclinepi.2005.10.015

27. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307–10. doi:10.1016/S0140-6736(86)90837-8

28. Terwee CB, Roorda LD, Knol DL, de Boer MR, de Vet HC. Linking measurement error to minimal important change of patient-reported outcomes. J Clin Epidemiol. 2009;62:1062–67. doi:10.1016/j.jclinepi.2008.10.011

29. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol. 2010;10:22. doi:10.1186/1471-2288-10-22

30. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2011 Jul 6. [Epub ahead of print]

31. Endicott J, Nee J. Endicott Work Productivity Scale (EWPS): a new measure to assess treatment effects. Psychopharmacol Bull. 1997;33:13–6.

32. Uguz S, Inanc BY, Yerlikaya EE, Aydin H. Reliability and validity of Turkish form of Endicott Work Productivity Scale. Turk Psikiyatri Derg. 2004;15:209–14.

33. Erickson SR, Guthrie S, Vanetten-Lee M, Himle J, Hoffman J, Santos SF, et al. Severity of anxiety and work-related outcomes of patients with anxiety disorders. Depress Anxiety. 2009;26:1165–71. doi:10.1002/da.20624

34. Sanderson K, Tilse E, Nicholson J, Oldenburg B, Graves N. Which presenteeism measures are more sensitive to depression and anxiety? J Affect Disord. 2007;101:65–74. doi:10.1016/j.jad.2006.10.024

35. Lam RW, Michalak EE, Yatham LN. A new clinical rating scale for work absence and productivity: validation in patients with major depressive disorder. BMC Psychiatry. 2009;9:78. doi:10.1186/1471-244X-9-78

36. Lerner D, Amick BC,III, Rogers WH, Malspeis S, Bungay K, Cynn D. The Work Limitations Questionnaire. Med Care. 2001;39:72–85. doi:10.1097/00005650-200101000-00009

37. Lerner D, Amick BC, Lee JC, Rooney T, Rogers WH, Chang H, et al. Relationship of Employee-Reported Work Limitations to Work Productivity. Med Care. 2003;41:649–59. doi:10.1097/01.MLR.0000062551.76504.A9

38. Koopman C, Pelletier KR, Murray JF, Sharda CE, Berger ML, Turpin RS, et al. Stanford presenteeism scale: health status and employee productivity. J Occup Environ Med. 2002;44:14–20. doi:10.1097/00043764-200201000-00004

39. Jette AM, Davies AR, Cleary PD, Calkins DR, Rubenstein LV, Fink A, et al. The Functional Status Questionnaire: reliability and validity when used in primary care. J Gen Intern Med. 1986;1:143–9. doi:10.1007/BF02602324

40. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993;78:98–104. doi:10.1037/0021-9010.78.1.98

Received for publication: 21 January 2011