Health Psychology

Master Thesis

Systematic review of patient-reported depression measures in rheumatoid arthritis

Sabine L. Kowoll

1st Supervisor: Dr. A.H. Oude Voshaar

2nd Supervisor: Dr. P.M. ten Klooster

Submitted to the Department of Behavioral Sciences, University of Twente,
in partial fulfilment of the requirements leading to the award of
Master in Health Psychology.

August 2018

UNIVERSITY OF TWENTE.

UNIVERSITY OF TWENTE.

Behavioral Sciences

Health Psychology

Master of Sciences

Systematic review of patient-reported depression measures

in rheumatoid arthritis

Abstract

Psychological well-being is often disturbed in physical chronic disease conditions like rheumatoid arthritis [RA]. The purpose of this study was to assess the extent to which patient-reported outcome measures for depression are valid and reliable in this specific population, to gather evidence on measurement properties, and to give recommendations concerning the applicability of valid depression measures for different purposes. A systematic and comprehensive search of literature on the development or psychometric evaluation of patient-reported depression measures in RA was done and the included studies were reviewed. Content validity was assessed through examination of relevance and comprehensiveness of the items, items were linked to DSM-5 criteria for major depression. Further measurement property analyses were conducted with the COSMIN checklist and corresponding criteria. The included studies concerned two depression measures, Beck's Depression Inventory [BDI] and the Center for Epidemiologic Studies Depression Scale [CES-D]. Evaluation of relevance and comprehensiveness revealed the BDI to be closer associated with DSM-5 criteria for major depression than the CES-D. The results of this review revealed that both measures consist of more than one factor and calculations of separate factor scores may be more informative than a total score. Besides, both measures are suspected to be contaminated by somatic items which may be caused by the rheumatic disease rather than by depressive symptomatology. Also, strong associations with measures of disability and anxiety were reported, limiting the measures' ability to assess specific depressive symptomatology. All these aspects and recommendations for the use of these measures for different purposes are described and discussed in detail.

Table of content

Introduction

Rheumatoid arthritis, one of the most severe and common types of arthritis, is a chronic autoimmune disease with unknown cause and an ongoing yet uncertain disease progression. It is characterized by inflammation of joints and accompanied by symptoms as pain, fatigue, and stiffness, as well as damage and deformity of joints and bones, and physical disability as a consequence of chronic inflammation (Tehlirian & Bathon, 2008). Nowadays, treatment of RA targets at physical symptom reduction, primarily to avoid inflammations in order to prevent (further) damage of joints and bones. According to a systematic review of incidence and prevalence studies of RA (Alamanos, Voulgari & Drosos, 2006), prevalence rates vary between 0.2% and 1.2%, and are notably higher for women. The evident overrepresentation of female RA patients is relevant to note, as depression is also more common in women (Angst, Gamma, Gastpar, Lépine, Mendlewicz & Tylee, 2002; Kuehner, 2003; Nolen-Hoeksema, 1990; Sonnenberg, Beekman, Deeg & van Tilburg, 2000; Weissman et al., 1996).

In addition to the physical symptoms, RA places an enormous burden on patients' health-related quality of life [HR-QOL]. Mental health and psychological well-being may be negatively affected, and RA patients often experience depression (Dominick, Ahern, Gold & Heller, 2004; Kosinski et al., 2002). Compared to the general population in Western, industrialized countries, RA patients are twice as likely to suffer from depression, with prevalence rates between 13% and 20% (Dickens & Creed, 2001), whereas depression rates in the general Dutch population range from 5% to 10% (de Graaf, ten Have & van Dorsselaer, 2010). Adding further evidence, a high proportion of RA patients report chronic and intermittent levels of depression, which are associated with worse functioning and poorer health (Morris, Yelin, Panopalis, Julian & Katz, 2011). Additionally, associations are found between depression and higher levels of pain, fatigue, and disease activity (Sheehy, Murphy & Barry, 2006; Wolfe & Michaud, 2009), as well as work disability (Löwe et al., 2004). The causal relationship between pain and low mood is assumed to work in both directions (Nagyova, Stewart, Macejova, van Dijk & van den Heuvel, 2005; Newman & Mulligan, 2000). Consequently, pain does not only negatively affect psychological well-being but is influenced by increased levels of depression, too (Dickens, McGowan, Clark-Carter & Creed, 2002). Further, depression in RA is associated with lower treatment compliance (Sheehy et al., 2006), as well as higher suicide risk (Timonen et al., 2003), and mortality (Ang, Choi, Kroenke & Wolfe, 2005).

In many cases, depression in RA remains unnoticed and, as a consequence, also untreated (Dickens et al., 2001). Hence, it is of crucial importance to screen RA patients for depression in clinical settings. In addition, screening is important in psychological interventions directed towards general aspects of RA, wherein patients with clinical depression need to be detected in order to exclude them from studies. Further, depression may be an important outcome in interventions targeted at the improvement of quality of life [QoL] and serve as primary or secondary outcome measure in interventions targeted at psychological wellbeing.

Whereas medical treatment of RA mainly aims at physical improvement of joint swellings and pain reduction, the psychological problems associated with the disease may be better treated with evidence-based psychological interventions. Indeed, different types of psychological interventions for the improvement of psychological functioning exist, e.g., for depression, pain, or coping. The effectiveness of such interventions has been examined in several reviews and meta-analyses. Conclusions drawn from a review (Astin, Beckner, Soeken, Hochberg & Berman, 2002) reveal that psychological interventions such as cognitive behavioral therapy, relaxation, and stress management may be effective complements to the conventional medical management of RA.

In contrast to physiological measures as heart rate and blood pressure, subjective concepts of mental health like depression may not be assessed objectively by means of direct measurement. Concerning clinical diagnoses of depression, assessment based on the criteria for a major depressive episode [MD] of the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders [DSM-5] (American Psychiatric Association, 2013) constitutes the gold standard. However, assessment of DSM-based diagnoses is not easily utilized in clinical settings and in the context of clinical studies. Consequently, self-reported questionnaires are often used to assess subjective facets of mental health from the patients' perspective.

Throughout the years, many patient-reported outcome measures [PRO] for depression have been developed, for example, the Beck Depression Inventory [BDI] and revised versions of it (Beck, Ward, Mendelson, Mock & Erbaugh, 1961; Beck, Rush, Shaw & Emery, 1979; Beck, Steer & Brown, 1996), the Center for Epidemiologic Studies Depression Scale [CES-D] (Radloff, 1977), the Hospital Anxiety and Depression Scale [HADS] (Zigmond & Snaith, 1983), and the Patient Health Questionnaire-9 [PHQ-9] (Spitzer, Kroenke, Williams & the Patient Health Questionnaire Primary Care Study Group, 1999), among others.

Due to the wayward character of these PROs, clear and unambiguous conclusions concerning a diagnosis of depression cannot be easily drawn. Although developed to assess the same concept, PROs differ in their number of items, question construction and wording, response format, and scoring. Indeed, an individuals' score on one PRO may differ strongly from those of another PRO; a patient may be rated as depressive by one measure, whereas another results in a score indicating no depression. In practice, cut-off scores, commonly developed from and for the general population, are used to draw conclusions from scores on a PRO. Hence, the results of PROs need to be interpreted with caution as conclusions drawn may not be valid for all populations. Consequently, application of PROs for depression in the context of RA requires verification of cut-off scores, and validation studies have to be conducted to assess the measurement properties of various PROs for different populations.

The most important psychometric properties for a PRO, used for screening or as an outcome measure, are reliability and validity. With measurement instruments that are neither reliable nor valid, screening for depression may result in wrong conclusions and misclassification of patients, and the effectiveness of psychological interventions may not be assessed adequately. Undoubtedly, there is a strong need for valid and reliable instruments. Therefore, studies assessing the measurement properties of PROs should be conducted as well as reviews summarizing these scientific efforts and their results in order to obtain evidence-based knowledge of the selection of adequate measurement instruments.

Striving to close a scientific gap, the purpose of this study is to systematically review studies investigating measurement properties of PROs for depression validated for RA patients. The measurement properties of the instruments will be analyzed and systematically judged according to DSM-5 criteria for major depressive episodes [MD] and criteria developed by the Consensus-based Standards for the selection of health Measurement Instruments [COSMIN] initiative (Mokkink et al., 2010a). The outcomes of this review should answer the question to what extent different PROs for depression are empirically supported for use in RA patients, as well as to point out aspects requiring scientific attention and further validation in future studies. Additionally, recommendations concerning the applicability of depression PROs as screening and outcome measure in RA will be made. Such findings are important in the collection of scientific knowledge as well as in clinical routine and practice.

Method

Study selection

A systematic and comprehensive literature search was performed with the intention to identify all relevant articles concerning the development or psychometric evaluation of PROs assessing depression validated for use in adult RA patients. In order to find all potentially eligible studies, a validated and sensitive search filter, developed by Terwee, Jansma, Riphagen, and de Vet (2009) was used. As specific types of keywords had to be included, the search strategy consisted of three different sets of independent searches which were merged into the final search string. The first search block concerned the concept to be measured, i.e., depression. The second block consisted of the population of interest, i.e., RA patients. The validated and sensitive search filter for the identification of studies on measurement properties of health-related PROs (Terwee et al., 2009) made up the third block. Eventually, these three search blocks were merged together to the fourth and final block that was applied to the Scopus (1975-2013) and PubMed (1973-2013) databases in February 2013. As two databases with varying search modules were used, the search string developed for use on PubMed had to be slightly adapted for use on Scopus. The precise search strings may be derived from the supplementary material (appendix 1.1. and 1.2.). More information about the validated search filter applied for the literature search is described in Terwee et al. (2009).

In order to select all relevant studies for further analysis, titles and abstracts of the articles were independently screened by two reviewers (Oude Voshaar & Kowoll). To be considered eligible, a number of inclusion criteria had to be fulfilled. Most importantly, the studies' main focus had to concern the development or psychometric evaluation of a PRO for depression in adults. PROs assessing depression as part of a more global psychological health status were not taken into account. Further, articles had to be published in English and studies must have assessed the original language version of a PRO. The application of PROs in languages and cultures other than those the measures were originally developed in requires adequate translation. Cross-cultural translation is a complex, iterative process wherein forward and backward translations with native speakers and professionals from the field should be made. Literal translations may lead to measures that are not culturally relevant or lack conceptual, semantic, operational, or item equivalence (Hewlett et al., 2016). Researchers and clinicians wanting to apply a PRO in their country are tempted to

translate it themselves without participation of native speakers, professionals, or the PRO developer. Precise reports of the translation procedure are rarely published. As a consequence, measurement properties of translated versions cannot be compared to those of the original version without firstly investigating the quality of the translation process and its' results. An appropriate translation procedure cannot be taken for granted, therefore studies examining translated versions of a PRO were excluded. The inclusion of RA patients in the study population was another essential criterion; i.e., analyses and results for this part of the study population must have been reported separately. In case of study populations with various disease groups and no separate analyses, the study population must have consisted of at least 50% RA patients. In addition, studies were excluded if analyses were reported for fewer than 50 patients, as the quality criteria for measurement properties applied in this study require at least 50 patients per analysis to be eligible for rating. Discrepancies in judgment of eligibility of studies were resolved by discussion and the final decision on the studies included in this review was made by consensus.

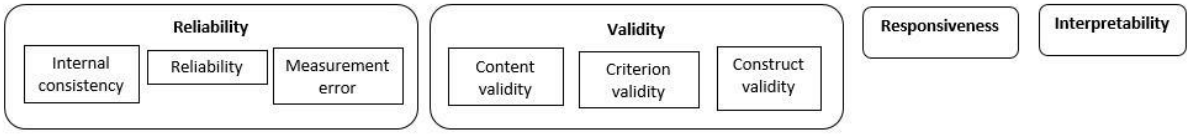
For all studies, information on which measurement properties were assessed for which PROs and on characteristics of the study population (sample size, mean age, percentage of RA patients and female participants) was extracted. Information on measurement properties was identified by use of the COSMIN checklist (Mokkink et al., 2010b), which was developed in a Delphi study of the COSMIN initiative, a multidisciplinary, international collaboration of experts in the field of health outcome measurement. For the development of the checklist, terminology, definitions, and taxonomy of measurement properties were discussed and agreed upon. The checklist contains standards for the evaluation of methodological qualities of studies on measurement properties of PROs. For all included studies, the checklist was scored independently by two reviewers (Oude Voshaar & Kowoll) according to the instructions described in the appendant COSMIN manual. Again, discrepancies in judgment were resolved by discussion.

Reliability (internal consistency), and criterion and construct validity were then rated according to quality criteria proposed for the COSMIN checklist (Terwee, Bot, de Boer, van der Windt, Knol, Dekker & Bouter, 2007). Content validity was rated using another approach described in the corresponding section below. An overview of the quality criteria for the measurement properties is presented in the supplemental material (appendix 2.1.).

Measurement properties

According to the COSMIN initiative, the taxonomy of measurement properties contains three main domains, namely validity, reliability, and responsiveness, with each of the domains consisting of one or more measurement properties (see figure 1, Mokkink et al., 2010b). The domain reliability incorporates three measurement properties, namely reliability, internal consistency, and measurement error. Comprised together, the measurement properties content validity, criterion validity, and construct validity constitute to the domain of validity. In turn, the domain responsiveness contains just one measurement property, also called responsiveness. The measurement properties of reliability and validity are further differentiated into aspects. Although not contained as a separate domain in the COSMIN taxonomy, interpretability is another important characteristic. All these measurement properties are assumed to be relevant and should be evaluated for HR-PROs.

Figure 1. The COSMIN taxonomy of measurement properties



Validity.

In general, validity is defined as the degree to which a scale measures what it intends to measure (McDowell, 2006). A joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education, defined validity as the evidence for inferences made about a test score, and agreed upon three types of evidence, namely construct-related, criterion-related, and content-related validity (Kaplan & Saccuzzo, 2009). These three types of validity are also represented in the COSMIN checklist. In research and clinical practice, DSM-V is often used as kind of a gold standard for the assessment of depression. According to the COSMIN initiative, there exist no gold standards for PROs except for shortened versions of a measure (Mokkink et al., 2010b) and this reasoning was followed for the purpose of this review.

Content validity.

Content validity refers to the degree to which the content of a measurement instrument adequately reflects the construct to be measured (Mokkink et al., 2010b). Thus, appraisal of content validity requires an evaluation of relevance and comprehensiveness of items. Relevance is judged by seeking answers to the questions whether all items refer to relevant aspects of the construct to be measured, whereas comprehensiveness refers to the degree to which the construct to be measured is covered by the items contained in a given PRO. A commonly agreed upon standard for the assessment of depression are the DSM criteria for MD (American Psychiatric Association, 2013). These criteria were used to evaluate content validity of the PROs in this study; relevance was rated positive if all items contained in a given PRO referred to DSM-5 criteria for MD, whereas comprehensiveness was rated positive if all DSM-5 criteria for MD were covered by the items in a PRO (Mokkink, 2010b). For this analysis, all items were linked to DSM-5 criteria for MD independently by two reviewers (Oude Voshaar & Kowoll); discrepancies in judgment were discussed until consensus was reached.

Construct validity.

Construct validity, defined as the degree to which scores of a measurement instrument are consistent with (theoretically derived) hypotheses (concerning the constructs to be measured) (Mokkink et al., 2010a), should be used to provide evidence of validity if criterion validity cannot be assessed due to the lack of a gold standard. These hypotheses may concern internal relationships, relationships to other instruments, or differences between relevant groups. In order to assess construct validity, predefined hypotheses have to be tested. For example, if measures assessing theoretically similar or identical constructs correlate highly with each other, evidence for convergent validity is demonstrated. In turn, discriminant validity is supported if theoretically different constructs are minimally correlated with each other. Without specific, predefined hypotheses, the risk of bias increases as it is easier to think up alternative explanations for low correlations than to conclude that an instrument may not be valid. Concerning hypothesis testing, the COSMIN initiative presents no standard for the number of hypotheses to be tested in a construct validity study. Nevertheless, the more hypotheses and the more specifically formulated, the more evidence is gathered for construct validity. Thus, hypotheses should precise the

direction and the magnitude of expected correlations. For a positive rating of construct validity, hypotheses have to be formulated a priori and at least 75% of the results have to be in accordance with these hypotheses in samples larger than 100.

Structural validity, referring to the degree to which scores on a HR-PRO are an adequate reflection of the dimensionality of the construct to be measured, is another aspect contributing evidence for construct validity. Systematical assessment of structural validity of constructs as depression is difficult as the underlying measurement model, which could be either reflective or formative, is rather equivocal than clearly evident. Neither the COSMIN initiative nor Terwee et al. (2007) present specific quality criteria for the assessment of structural validity. A definite judgment of quality is only possible if confirmatory factor analysis[CFA] was applied. Thus, it was decided to give a positive rating if CFA was applied in a sample with a minimal size of the tenfold number of items, with at least a hundred participants. Also, sufficient information on model fit must be presented and values of model fit must be acceptable for the supposed underlying measurement model of the PRO.

Criterion validity.

Criterion validity is defined as the degree to which scores on a particular measure are an adequate reflection of a gold standard. In contrast to physiological measures, constructs as depression cannot be measured objectively. According to the COSMIN initiative, there exist no gold standards for HR-PROs except for comparisons of shortened versions to their original long version. Also, there is no commonly agreed upon standard for the assessment of depression for various purposes. It was decided to follow this reasoning; only original versions were considered as gold standard in the assessment of criterion validity. For a positive rating of criterion validity, convincing arguments must have been presented for the gold standard, that is, a shortened version of a PRO must have been compared to the original version, and correlation with this gold standard must have been ≥ 0.70.

Additionally, an overview of the degree to which the discriminative ability of the PROs has been compared to ‘gold-standard’ diagnostic criteria (e.g. clinical interview using MID, SCID-I, etc.) will be presented in order to summarize evidence concerning the PROs’ ability to be used for screening for the presence of MD. There are numerous methods to assess a questionnaires ability to discriminate between depressed and non-depressed patient. As an overall measure of discriminative ability, the area under the

receiver operating curve [AUC] is most commonly used. Discriminative ability was judged to be acceptable if AUC values were ≥ 0.80. Besides, the sensitivity and specificity of specific cut-off points will be presented.

Reliability.

Overall, reliability concerns the extent to which scores for patients who have not changed on the construct to be measured are the same for repeated measurement under certain conditions. More precisely, reliability refers to the extent to which measurement is free from measurement error.

Internal consistency.

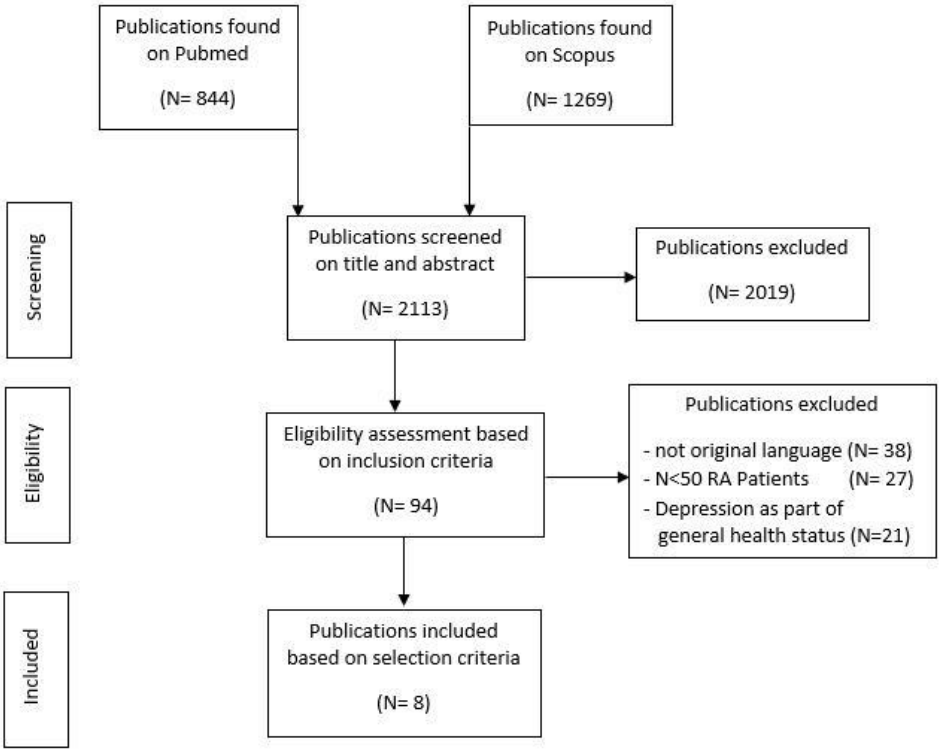
Assessment of internal consistency is only relevant if the items together form a reflective model. In a reflective model, the variance of scores is assumed to be caused by the measured trait + measurement error. Therefore, items should be highly inter-correlated. Nevertheless, too high correlations among the items may indicate redundant content. Often, it is not explicitly stated whether a measures' construct is based on a reflective or formative model. Another prerequisite for internal consistency statistics to get an interpretable meaning is that the scale needs to be unidimensional. For a positive rating, factor analysis should be performed and indicate homogenous scales in sufficiently large samples (N = 7* #items and ≥ 100) and Cronbach's α should be calculated for each dimension and range from 0.70 to 0.95.

Results

Study selection

The systematic literature search resulted in a total number of 2113 hits (Pubmed: 844, Scopus: 1269). Initial screening of titles and abstracts led to the exclusion of 2019 articles. The remaining 94 articles were further examined by two reviewers (Oude Voshaar & Kowoll) to judge their eligibility based on the inclusion criteria. Out of these articles, 38 did not assess the original language version of a PRO, 27 articles were excluded due to study populations with less than 50 RA patients, and 21 were excluded because of not assessing depression primarily but as part of a more general health status assessment. Eventually, eight studies were identified that examined the psychometric properties of PROs concerning depression and met all inclusion criteria. In the included studies, the measurement properties of only two different PROs were examined, namely the BDI and the CES-D. Study selection is outlined in figure 2., an overview of the measurement properties assessed as well as sample information is presented in the additional material (appendix 3.1. and 3.2.). On the whole, two studies concerned the BDI and its construct validity, while six studies assessed diverse measurement properties of the CES-D.

Figure 2. Study selection procedure



The BDI, a 21-item self-report instrument, intends to measure depression symptoms and severity through items concerning cognitive, affective, somatic, and vegetative symptoms of depression (Beck et al., 1961). After various revisions of item wording, substantial changes with regard to content were made in BDI II (Beck et al., 1996) which was developed to correspond to DSM-4 criteria (American Psychiatric Association, 2000) for major depressive disorder [MDD]. Responses to the BDI refer to the timeframe of the last two weeks and are scored on a 4-point scale indicating degrees of depression severity from 0 (“not at all”) to 3 (“extreme form of symptom”), with a total score range from 0 to 63.

The purpose of the CES-D is to measure current levels of depressive symptoms. The original version contains 20 items assessing perceived mood and functioning over the past week. According to Radloff (1977), four factors are present: depressed affect [DA], positive affect [PA], somatic complaints and retarded activity [SC], and interpersonal relationships [IP]. Multiple shortened versions were developed for use with various populations but these are found to overestimate patients with chronic diseases like RA as being depressed (Zauszniewski & Bekhet, 2009). For the original 20-item version, responses are scored on a 4-point scale from 0 (“rarely/ none of the time”) to 3 (“most/ all of the time”), total score ranges from 0 to 60.

Measurement properties

The following sections describe the studies included in this review, the statistical methods applied and their results, as well as the measurement property quality rating assigned according to the criteria proposed by the COSMIN initiative. Table 3.1. presents an overview of the quality ratings.

Table 3.1. Quality rating of measurement properties

	Relevance	Comprehen-siveness	Construct validity	Structural validity	Criterion validity	Internal consistency
BDI	? (81%)	? (8 criteria)	? [1,6]	? [1,6]	0	0
CES-D	- (45%)	- (5 criteria)	? [5]	+ [7,8]	? [5]	? [5]
CES-D-SF	-	-	0	0	? [2] + [4]	0

+ = good measurement properties with adequate methodological quality; ? = indeterminate quality of measurement properties because of inadequate methodological quality; - = poor measurement properties despite adequate methodological quality; 0 = no information found; [1] = Hagglund; [2] = Martens 2003; [3] = Martens 2005; [4] = Martens 2006; [5] = McQuillan; [6] = Peck; [7] = Rhee; [8] = Sheehan.

Validity.

Content validity.

For the assessment of content validity, all BDI and CES-D items were independently linked to DSM-5 criteria for MD by two reviewers (Oude Voshaar & Kowoll). Agreement reached 80% for the CES-D and 95% for the BDI, resulting in an overall agreement of 87,5%. Initially, five items were not linked in accord; consensus was reached through discussion. With the results of the linking procedure at hand, relevance and comprehensiveness of the items were evaluated.

To judge relevance, it was examined whether all items refer to DSM-5 criteria for MD. The extent to which the items could be linked to these criteria varied; whereas 17 out of 21 BDI items were linked unambiguously, only nine out of twenty CES-D items referred to DSM-5 criteria for MD. Thus, four BDI items and eleven CES-D items were considered irrelevant; these items refer to hopelessness, irritability, avolition, hypochondria, anxiety, social withdrawal, loneliness, and feeling unloved. The results of the linking procedure for each item are presented in the supplemental material (appendix 3.3. & 3.4.).

Comprehensiveness was judged by analyzing the extent to which the nine DSM-5 criteria for MD were covered by the items of a PRO. All DSM-5 criteria for MD except of psychomotor agitation/retardation were covered by BDI items. Concerning the CES-D, four criteria, i.e. anhedonia, psychomotor agitation/retardation, fatigue/loss of energy, and suicidality, were not covered by the items. An overview of the DSM-5 criteria coverage of the PROs is presented in table 3.2. The results also reveal that the nine DSM-5 criteria for MD are unevenly covered by the items. A majority of six BDI items refer to worthlessness/guilt, whereas the other criteria are covered by one to three items, only. Concerning the CES-D, it was found that four items refer to dysphoria, although only one or two items refer to the other criteria.

Applying the quality criteria presented in the method section, the following ratings were given. BDI was rated indeterminate for both, relevance and comprehensiveness, as 81% of the items were linked to DSM-5 criteria for MD and eight out of nine criteria were covered by the items. In contrast, relevance and comprehensiveness were rated negative for the CES-D because only 45% of the items were linked to DSM-5 criteria for MD and only five DMS-5 criteria were covered by the items.

Adding further information on the content of BDI, item disease relevance was rated for by fifteen rheumatologists in one of the reviewed studies (Peck et al., 1989). According to these ratings, eight BDI items refer to RA-related symptomatology.

Table 3.2. Number of items associated with DSM-V criteria for MD

DSM-V criterion for MD	BDI	CES-D
Dysphoria	2	4
Anhedonia	3	-
Changes in appetite/weight	2	1
Insomnia/hypersomnia	1	1
Psychomotor agitation/retardation	-	-
Fatigue/loss of energy	1	-
Worthlessness/guilt	6	2
Cognitive difficulties	1	1
Suicidality	1	-

Construct validity.

Overall, three studies gathered direct evidence of construct validity, out of which two concerned the BDI (Hagglund, Roth, Haley & Alarcón, 1989; Peck, Smith, Ward & Milano, 1989), and one the CES-D (McQuillan, Fifield, Sheehan, Reisine, Tennen, Hesselbrock & Rothfield, 2003). Further, both BDI studies and two other CES-D studies (Rhee, Petroski, Parker, Smarr, Wright, Multon, Buchholz & Komatireddy, 1999; Sheehan, Fifield, Reisine & Tennen, 1995) assessed structural validity, an aspect of construct validity according to the COSMIN taxonomy. For both BDI studies, the measurement properties assessed were rated as indeterminate due to inadequate methodological quality. Concerning construct validity, both studies tested predefined hypotheses including the direction of correlations of the BDI with other measures but failed to precisely formulate the magnitude of these expected correlations and could thus not be rated positively. With 52 participants, Hagglund et al. (1989) failed to fulfil the criterion for sample size.

In both studies, examination of convergent and divergent validity, also referred to as discriminant validity, was carried out to gather evidence of construct validity. Therefore, hypotheses concerning the direction of expected relations of the BDI with commonly used measures of affective distress, i.e., the Arthritis Helplessness Index [AHI], the state-trait-anxiety inventory [STAI], the depression and anxiety scales of the Arthritis Impact Measurement Scales [AIMS] (Hagglund et al., 1989), and with self-reported disability (Health Assessment Questionnaire [HAQ] disability scale) as well as observation-based disability and depression (interview-based Hamilton Rating Scale for Depression [HRSD]) were formulated a priori (Peck et al., 1989). Hagglund et al. (1989) expected stronger positive correlations between the BDI and the depression scale of the AIMS and AHI than with any of the anxiety measures. Peck et al. (1989) expected the strongest relationships between BDI and HRSD, whereas correlations between the depression and disability measures were predicted to be smaller. Pairwise correlations among the scales (Hagglund et al., 1989) and Pearson correlation coefficients (Peck et al., 1989) were calculated from data of 52 and 107 RA patients, respectively.

Overall, the results of correlational analyses confirmed the hypothesized relationships between the BDI and the other measures. According to Hagglund et al. (1989), the depression and anxiety measures were all significantly correlated, with correlations ranging from 0.61 – 0.82. BDI correlated most strongly with the depression scale of the AIMS ($r = 0.82$), followed by a slightly lower correlation with the TAI ($r = 0.78$). The

lowest correlations for the BDI were found with the AHI (r = 0.56) and the anxiety scale of the AIMS (r = 0.62). Peck et al. (1989) found confirmation for the expected positive correlation between the two depression measures (r = 0.69). Significant correlations with varying magnitude were also found across all pairs of disability and depression measures. Whereas correlations between disability measures and HRSD were quite small (r = 0.17 – 0.25), the BDI correlated stronger with disability measures (r = 0.31 – 0.50), indicating either artifacts associated with the self-report method or contamination through somatic items (Peck et al., 1989).

Using another approach, Hagglund et al. (1989) assessed the dimensionality of the BDI through CFA, wherein a unidimensional distress factor was compared to a two-factor-model positing separate depression and anxiety factors. Factor loadings and inter-correlation estimates were obtained with a maximum likelihood estimation technique; model fit was examined by chi-square values [x^2] and the goodness-of-fit index [GFI]. It was hypothesized that the two-factor-model would best explain the data if the scales have high levels of both, convergent and divergent validity. In case of good convergent but poor divergent validity, the one-factor-model would explain the data better. Results of CFA revealed that the one-factor model fit the data fairly well although both x^2 and GFI indicated some room for improvement. Factor loadings ranged from 0.606 for AHI to 0.895 for BDI. The two-factor model was found to fit significantly better than the one-factor model (x^2 difference = 6.36, df = 1, $p < 0.05$), with BDI loading 0.94 on the depression factor. Nevertheless, the two factors correlated highly (r = 0.90), suggesting little or no conceptual uniqueness between the two constructs. Analysis of a three-factor model adding an AIMS factor revealed that this model provided adequate fit but the correlations of 0.90 between the two factors remained, confirming the finding that there is virtually no separation between the constructs of depression and anxiety on these measures. Table 3.3. presents the factor labels and loadings resulting from CFA of the BDI, as well as the item allocation (Hagglund et al., 1989); Peck et al. (1989) have not reported precise item allocation and factor loadings.

Further, structural validity was assessed in both studies but no positive rating could be assigned. In both studies, sample sizes were too small for positive ratings of analyses and item loadings on the factors were not reported. Peck et al. (1989) also failed to apply CFA as adequate statistical method in their analyses. Thus, the quality of structural validity is rated as indeterminate due to doubtful design and/ or method for both studies.

Table 3.3. BDI item allocation, factor labels and loadings

#	Item	Factor label	Factor loading
1	Sadness	DM	0.70
2	Future pessimism	DM	0.67
3	Failure	DM	0.76
4	Enjoy	SC	0.63
5	Guilt	DM	0.68
6	Punishment	DM	0.51
7	Disappointed	DM	0.66
8	Blame	DM	0.66
9	Suicide	DM	0.72
10	Cry	DM	0.58
11	Irritated	*	*
12	Interest in other people	DM	0.49
13	Decision making	*	*
14	Appearance	DM	0.51
15	Work	SC	0.76
16	Sleep	SC	0.43
17	Tired	SC	0.62
18	Appetite	SC	0.60
19	Weight	SC	0.49
20	Worry	SC	0.58
21	Interest in sex	SC	0.45

DM = dysphoric mood, SC = somatic complaints, * = loaded on both factors.

To examine the assumption that the BDI may be contaminated by disease-related items which may rather reflect symptoms associated with RA rather than depression in this specific population, Peck et al. (1989) subjected items to principal components analysis [PCA], applied varimax rotation to factors with an eigenvalue > 1.0 and individual items were considered to load on a given factor if the loading value was > 0.40 on only one factor (Peck et al., 1989). Results of PCA demonstrated the presence of two components, only two items could not be clearly assigned to one of these two components which were

labeled as “dysphoric mood” [DM] and “somatic complaints” [SC]. Six out of eight SC items and only two out of eleven DM items were identified as reflecting RA by rheumatologists. Both of these BDI components were significantly correlated with HRSD scores, with a stronger correlation for the DM component. Also, this component correlated stronger with HRSD scores than with all disability measures, the SC factor did not. In turn, the SC component correlated stronger with disability measures than the DM component. These results support evidence of convergent and divergent validity. Still, the BDI reflected some somatic contamination and use of a total score is thus likely to cause inaccurate results in RA populations. Therefore, a DM subcomponent, demonstrating good convergent and divergent validity, may be a more valid measure of depression in RA as the SC factor is likely to produce misleading results if interpreted as a measure of depression in this population.

Summed up, the results from correlational and factorial analyses of both BDI studies indicate adequate convergent validity but poorer discriminant validity due to the high correlation between the factors, limiting the ability and utility of these measures to effectively distinguish among separate problems with depression and anxiety in RA in clinical and research settings.

CES-D construct validity was examined in one study (McQuillan et al., 2003) for which no positive measurement property quality rating could be assigned due to the absence of properly formulated hypotheses, including the direction and magnitude of expected correlations between measures. Still, requirements for study design and sample size were fulfilled. The study assessed a sample of 415 RA patients and evaluated the discriminant validity of the CES-D, the Positive and Negative Affect Schedule [PANAS], and the Endler Multidimensional Anxiety Scale [EMAS], a measure of state anxiety, specifically designed to distinguish anxiety from depression. These scales` ability to discriminate between a disorder, no disorder, as well as between types of disorder (MD, Generalized Anxiety Disorder [GAD], or comorbid disorder [CD]) was assessed. Analyses contained bivariate correlations among full- and subscales as well as analysis of variance (ANOVA) tests of the differences in mean scale scores by affective disorder. The results of these analyses for the entire sample revealed adequate correlations between the CES-D subscales (all > 0.60). Also, each of the four subscales had a strong positive correlation with the full CES-D (0.80 – 0.93), demonstrating good convergent validity for the subscales. Correlations between the CES-D subscales, the EMAS subscales, and both of the PANAS

subscales indicate a limited ability to discriminate between depression and anxiety as some correlations between the depression and anxiety subscales were quite high. These positive correlations between the CES-D and anxiety subscales indicate that both scales tap negative affect. The overall pattern of correlations among participants with a diagnosis of affective disorder were similar to those of the full sample, convergent validity was indicated to a higher extent than discriminant validity.

Structural validity of the CES-D was examined in two studies (Sheehan et al., 1995; Rhee et al., 1999), which both received a positive rating of measurement property quality as criteria concerning sample size and statistical methods, i.e., to conduct CFA, were met and values of model fit were acceptable and appropriately reported. In both studies, adequate descriptions were given concerning sample characteristics and study settings, thereby improving the generalizability of results to RA populations.

Sheehan et al. (1995) compared four alternative measurement structures; a single-factor model positing one underlying variable, a three-factor model with PA and DA representing two ends of a single underlying affect dimension, Radloff's four-factor model (Radloff, 1977) to examine if the CES-D differentiates between PA and DA, as well as a second-order factor model positing a single second-order factor underlying the four-factor model. The same models and an additional three-factor model consisting of DA, PA, and IP were tested by Rhee et al. (1999). Here, all three- and four-factor models were analyzed with Radloff item allocation and the item allocation of Sheehan et al. (1995). In both studies, the best fitting models were cross-validated in two follow-up assessments to determine their temporal stability, an essential quality if scores based upon these structures are used to monitor change over time. Results revealed that the four-factor models demonstrated superiority over the single-factor and three-factor models and were statistically comparable with the second-order-factor models in both studies, indicating that the correlations among the four factors can be explained by a single second-order factor, i.e. depression. Fit indices for the multiple models examined in both studies are presented in table 3.4.

Temporal cross validation over two additional time points revealed that factor structure and loadings were stable over time in both studies. Although generally confirming the results of Sheehan et al. (1995), Rhee et al. (1999) found the item allocation of Radloff (1977) to be superior to those of Sheehan. Item allocation to the factors is presented in table 3.5., the only differences concerned the items *failure* and *fearful* which loaded on the factor IP in Sheehan et al. (1995). In contrast, the results of Rhee et al. (1999) confirm the item allocation of Radloff (1977), where these two items belong to the DA factor.

Table 3.4. Fit indices of CES-D factor models

	x^2	*df*	x^2 /*df*	RMSEA	AIC	FI*
Sheehan						
One-factor model	702	170	4.13	0.065	782	0.926
Three-factor model (DA+PA, SD, IP)	541	167	3.24	0.055	627	0.948
Four-factor model (DA, PA, SD, IP)	247	164	1.51	0.026	339	0.988
Second-order four-factor model	253	166	1.52	0.027	340	0.988
Four-factor correlated error	148	160	0.93	0.000	248	1.000
Second-order correlated error	154	162	0.95	0.000	250	0.997
Rhee						
One-factor model [R]	621	170	3.7	0.08	625	0.86
Three-factor model (DA+PA,SV,IP) [R]	450	167	2.7	0.07	356	0.90
Three-factor model (DA+PA,SV,IP) [S]	495	167	3.0	0.07	530	0.89
Three-factor model (DA+SV,PA,IP) [R]	408	167	2.4	0.06	289	0.91
Three-factor model (DA+SV,PA,IP) [S]	441	167	2.6	0.07	341	0.90
Four-factor model (DA, PA,SV,IP) [R]	299	164	1.8	0.05	131	0.94
Four-factor model (DA, PA,SV,IP) [S]	330	164	2.0	0.06	180	0.93
Second-order four-factor model [R]	305	166	1.8	0.05	135	0.94
Second-order four-factor model [S]	340	166	2.0	0.06	191	0.93

* = Fit indices: Sheehan et al. calculated CFI, while Rhee et al. calculated GFI; [R] = item allocation according to Radloff; [S] = item allocation according to Sheehan; DA = depressed affect; IP = interpersonal relations; PA = positive affect; SV = somatic / vegetative.

Although no direct evidence of criterion contamination was found in these two studies, the differences in item allocation raise questions concerning the content of the factors and potential contamination through items relating to symptoms of RA. Therefore, the authors (Sheehan et al., 1995; Rhee et al., 1999) conclude that use of a single summary score is clearly not the most informative in RA populations; rather one may compute separate factor scores and should be aware of potential criterion contamination in the SD factor.

Table 3.5. CES-D item allocation, factor labels and loadings

#	Item	Rhee et al. (1999)				Sheehan et al. (1995)			
		Factor-label	Factor loading*			Factor-label	Factor loading*		
			T1	T2	T3		T1	T2	T3
1	Bothered	SD	0.57	0.48	0.45	SD	0.67	0.77	0.76
2	Eating	SD	0.45	0.39	0.41	SD	0.56	0.77	0.76
3	Blues	DA	0.77	0.69	0.65	DA	0.88	0.90	0.90
4	Good	PA	0.50	0.40	0.43	PA	0.54	0.54	0.66
5	Mind	SD	0.51	0.54	0.49	SD	0.69	0.75	0.74
6	Depressed	DA	0.85	0.75	0.77	DA	0.94	0.92	0.94
7	Effort	SD	0.64	0.64	0.63	SD	0.73	0.78	0.79
8	Hopeful	PA	0.53	0.50	0.55	PA	0.69	0.70	0.75
9	Failure	DA	0.41	0.53	0.54	IP	0.88	0.84	0.83
10	Fearful	DA	0.46	0.47	0.57	IP	0.75	0.79	0.85
11	Sleep	SD	0.52	0.51	0.47	SD	0.53	0.54	0.59
12	Happy	PA	0.83	0.72	0.81	PA	0.88	0.90	0.93
13	Talk less	SD	0.47	0.57	0.62	SD	0.71	0.73	0.74
14	Lonely	DA	0.65	0.68	0.65	DA	0.76	0.81	0.85
15	Unfriendly	IP	0.61	0.56	0.42	IP	0.59	0.68	0.60
16	Enjoy life	PA	0.69	0.69	0.79	PA	0.87	0.90	0.91
17	Cry	DA	0.55	0.53	0.55	DA	0.79	0.80	0.85
18	Sad	DA	0.74	0.76	0.78	DA	0.86	0.88	0.91
19	Dislike	IP	0.63	0.80	0.79	IP	0.65	0.78	0.82
20	get going	SD	0.61	0.57	0.63	SD	0.71	0.73	0.75

DA = depressed affect; IP = interpersonal relations; PA = positive affect; SD = somatic disturbance (Hyun Rhee referred to this factor as ‘somatic/vegetative’); * = standardized factor loadings from correlated four-factor model with Radloff item allocation for study of Rhee, parameter estimates for four-factor model in Sheehan’s study.

Criterion validity.

Three studies assessed the criterion validity of the original CES-D (McQuillan et al., 2003) or modified, shortened versions (Martens, Parker, Smarr, Hewett, Slaughter & Walker, 2003; Martens, Parker, Smarr, Hewett, Slaughter & Walker, 2006).

McQuillan et al. (2003) received an indeterminate measurement property quality rating as various measures assessing depression and/ or anxiety were compared with each other out of which no one can be regarded as a gold standard like in the case of comparison of shortened versions with original scales. In this study, previous research findings of potential criterion contamination were further investigated; the CES-D, PANAS, and EMAS were compared with each other in terms of sensitivity and specificity, it was assessed whether somatic CES-D items artificially inflate scores, and evidence for an optimal cut off score in RA populations was gathered. The combined sensitivity and specificity of the CES-D with and without somatic items was compared using receiver operator characteristic [ROC] curves. Scores on CES-D, PANAS, and EMAS were compared to diagnostic criteria of MD, GAD, and CD; current and lifetime psychiatric diagnoses of MD, GAD, and CD were obtained using the Semi-Structured Assessment for the Genetics of Alcoholism [SSAGA], which is based on DSM-4. According to SSAGA scores, 9% of the sample was affected by an affective disorder. For analyses of discriminative ability, CD participants (N = 27) were eliminated as they cannot be placed in either group. The degree to which somatic items artificially inflate CES-D total score in RA was examined through comparison of a shortened version without somatic items (CES-Dnoso) with the original scale. Results of statistical examinations revealed that CES-Dnoso scores (mean = 10.17) were lower than those of the full scale (mean = 12.23), suggesting some criterion contamination. Nonetheless, magnitude of the difference in mean scores was small and the two scales were almost perfectly correlated (r = 0.99). Thus, the somatic items explained less than 3% of the original CES-D score (coefficient R^2 = 0.97). To determine optimal cut-off scores in RA, rates of true positives and false positives for various cut off scores were calculated. It was found that, compared with 16, only one true case was missed when 19 was used as a cut-off score but there were 22 more false positives with a cutoff score of 16. The authors conclude that all three measures have high combined sensitivity and specificity as measures of affective disorder among RA patients. Thus, it is possible to detect affective disorder in RA patients with the CES-D, which identified high levels of depression and anxiety equally well. Nevertheless, the CES-D was

not able to differentiate between MD and GAD, neither were PANAS or EMAS. ROC analyses further revealed that the CES-D had a significantly higher AUC than the other scales, indicating a better ability to differentiate between those with and without an affective disorder. No significant differences between AUC for the full CES-D and the shortened version without somatic items were found.

In another CES-D study, Martens et al. (2003) assessed the scales` ability to identify confirmed cases of MD and evaluated various cut-off scores for the full CES-D and a previously suggested modified version (Santor & Coyne, 1997) with nine items. Secondary analyses of data from 457 RA patients, out of which 91 met criteria for MD, were performed. It was hypothesized that a cutoff score from the modified CES-D would provide greater overall efficiency than the full-scale cutoff scores of 16 and 19. The study included an exploratory and a confirmatory phase, with sample sizes of 160 and 52 RA patients, respectively.

The authors conducted exploratory analyses to test various cutoff scores for the original CES-D and a modified version, which was scored dichotomously (Martens et al., 2003). Sensitivity, specificity, PPV, and NPV were calculated and compared for full scale cutoff scores of 16 and 19, and modified cutoff scores ranging from 3 to 8. The results of these analyses revealed that a full-scale cutoff score of 19 was more efficient in identifying cases of MD than 16 but also questionable, especially in terms of specificity and PPV. Compared to 16, a cutoff score of 19 resulted in a 10 points lower sensitivity value. Against expectations, the modified CES-D was less efficient in identifying cases of MD. The most efficient cutoff score for this version was 6, but none of the modified cutoff scores was as efficient as the full-scale cutoff score of 19.

Confirmatory analyses were conducted with data of 52 participants to replicate the results of the exploratory phase. Again, sensitivity, specificity, PPV, and NPV were calculated, but only for the most efficient cutoff scores, i.e., full scale 19 and modified 6. The results generally confirmed the findings of the first phase, the full-scale cutoff score of 19 was superior to the modified cutoff score of 6. Also, results for the modified cutoff score of 6 were similar in both phases. In the second phase, the full-scale cutoff score of 19 yielded even more efficient results (higher sensitivity, specificity, PPV, and NPV). Nevertheless, a sensitivity value of 0.86 for a full-scale cutoff score of 19 still indicates that 14% with MD were misclassified.

In these analyses, all participants had CES-D scores higher than 10. To test the established cutoff scores with a wider range of CES-D scores, a group of RA patients (N = 245) who never reported CES-D scores higher than 10 was added and additional analyses were conducted to address this limitation. Compared to a mean CES-D score of 30.1 (SD = 11.0) for participants diagnosed with MD in the first two phases, the mean CES-D score for the additional sample was 3.4 (SD = 3.0). The procedures from both phases were replicated with this additional sample. Overall, the results were consistent with previous study findings, a full CES-D cutoff score of 19 performed better in terms of identifying cases of MD than the modified cutoff scores. Summed up, the study demonstrated that the modified CES-D was less efficient in classifying cases of MD than the full CES-D. Further, a full-scale cutoff score of 19 provided greater overall efficiency than any of the cutoff scores of the modified version. Still, a cutoff score of 16 had higher sensitivity values than 19. Albeit being potentially useful as a screening tool, caution in decision making based on CES-D scores alone is advised as even the most efficient cutoff score resulted in patients being misclassified.

In a subsequent study, Martens et al. (2006) aimed to develop a CES-D short-form version for the identification of persons with MD within RA. The development of the modified CES-D (Santor et al., 1997), which Martens et al. (2003) examined in their previous study, was based on comparisons of responses on each CES-D item between a group of primary care patients with the diagnosis of MD, and a group of patients without a diagnosis of MD. Only items that revealed a large difference in symptom severity between the two groups were retained for the shortened version. According to Santor et al. (1997), cutoff scores from the shortened CES-D scale were more efficient than full-scale cutoff scores for identifying patients with MD in a primary care sample. This finding could not be replicated in a RA sample (Martens, 2003). Following the Santor approach of item selection, a shortened CES-D was developed with an optimized methodology for RA samples and multiple cutoff scores were examined. Analyses were based on existing longitudinal data from 337 RA patients out of which 46 met criteria for MD. Sensitivity, specificity, PPV, and NPV were calculated and compared for full-scale CES–D cutoff scores 16 and 19, as well as for multiple cutoff scores derived from the modified CES–D. Although traditionally scored on a 4-point scale (0 - 3), the scoring method was modified in this study and items were scored dichotomously (0 = “0”; 1-3 = “1”).

From the results of the scale development phase, nine items were selected for inclusion in the modified CES-D. Efficiency calculations indicated that a modified CES–D cutoff score of 5 was the most efficient short-form score (sensitivity = .96, specificity = .81, PPV = .44, NPV = .99) and generally as efficient as the more commonly used full-scale cutoff score of 16 for classifying participants with MD within RA. Use of a modified cutoff score of 4 yielded a value of 1.00 for sensitivity, i.e., all participants with MD were correctly classified. Although being superior to CES-D 16 in terms of sensitivity, values of specificity were slightly lower for the shortened cutoff score of 5. Further, ROC curves were generated for the original and modified CES-D to compare their efficiency. Overall efficiency of the modified CES-D (AUC = .94) was found to be equivalent to the original version (AUC = .95). Taken together, a cutoff score of 5 from the modified CES-D was generally as efficient as the more commonly used full-scale cutoff score of 16 for classifying participants with MD within RA. An overview of all efficiency values reported in the described studies is added in the supplemental material (appendix 3.5.).

Reliability.

Internal consistency.

For the assessment of internal consistency, McQuillan et al. (2003) received an indeterminate rating for the quality of this measurement property. The reason for this rating is that no factor analysis was conducted. Still, the authors report that all of the screening scales had adequate alpha reliabilities. More precisely, reliability coefficients for the subscales of the CES-D ranged from .71 for the interpersonal dimension, .83 for both, the somatic and positive affect dimension, to .88 for the depressive affect dimension. Further, none of the studies included in this review concerned issues of reliability.

Discussion

Taken the results of this review together, an overview of depression measures validated for RA and evidence on their measurement properties can be presented. Out of a large number of depression measures nowadays available, only two are validated in their original language for RA populations, i.e., the BDI and the CES-D. Overall, the CES-D received more attention in validation studies in RA populations than the BDI. Whereas six studies concerned the CES-D, only two were dedicated to investigate the BDI. Further, the scope of the investigations of the BDI was limited to the assessment of construct and structural validity. Validation studies of the CES-D did not only examine its' construct validity, but also concerned criterion validity.

The quality ratings of the measurement properties of the BDI and the CES-D should give an overview of the scientific evidence for the use of these PROs in RA populations. For these ratings, not only the results of the validation studies must have been promising, study design and population also must have been adequate for a positive rating. For the BDI, construct and structural validity was rated as indeterminate due to unfulfilled study design requirement in both studies (Hagglund et al., 1989; Peck et al., 1989). Nevertheless, the results of these studies still add knowledge to the usability of the BDI in RA but have to be interpreted with caution as the statistical methodology was not appropriate to receive a positive rating. More precisely, sample sizes were too small and CFA was not applied. Due to the lack of specific hypotheses, the CES-D was also rated as indeterminate for construct validity. Applying appropriate study design and statistical methods and reporting acceptable values of model fit, two studies (Rhee et al., 1999; Sheehan et al., 1995) added evidence for the structural validity of the CES-D and thus lead to a positive rating. Further, a shortened CES-D received a positive rating for criterion validity (Martens et al., 2006).

Concerning the results of the linking procedure for the assessment of content validity of the BDI and the CES-D, some noteworthy results were found. Neither relevance nor comprehensiveness were rated positively for any of the scales. For BDI, both aspects were rated as indeterminate, the CES-D received a negative rating of relevance and comprehensiveness. Referring to comprehensiveness, it is noteworthy that eight out of nine DSM-5 criteria for MD were covered by the BDI, whereas CES-D items could be linked to five criteria, only. Both PROs lack items on psychomotor agitation/ retardation, CES-D further does not cover anhedonia, fatigue/ loss of energy, and suicidality. A remarkable number of six BDI items were linked to the criterion of worthlessness/ guilt, whereas the

other criteria were covered by one to three items each. Concerning the CES-D, dysphoria was more extensively covered by the items than the other criteria.

These ratings of content validity mirror the coverage of DSM-5-MD criteria of the items, i.e., an indeterminate or negative rating does not disqualify any of the measures as a valid measure of depression. Rather, both PROs include items which tackle topics not included as criteria in the DSM-5. Less strict criteria for the quality rating of content validity might have resulted in more positive ratings. For example, items concerning hopelessness, irritability, and avolition were included in both PROs and could not be linked to the DSM-5-criteria for MD. In addition, one item of the BDI covers hypochondria and the CES-D also contains not linkable items referring to anxiety, social withdrawal, loneliness and feeling unloved. These items may not be deemed irrelevant per se only because they could not be linked to the DSM-5 criteria for MD. Through the inclusion of items covering complaints that are not part of the DSM-5 criteria for MD, a broader picture of the symptomatology of the patient may be given. Still, items which deviate from the content of the DSM-5 criteria for MD, e.g. items concerning anxiety, may also limit the measures ability to differentiate between depression and anxiety.

Implications for use as screening measure

Researchers and clinicians who seek to use PROs to screen for depression are faced with the question which measure to use to obtain valid and reliable results. Good criterion and construct validity would be supportive of a measures' utility for the purpose of screening. Based on the findings of this review, one may point out that there is more evidence available for the utility of the CES-D as screener for depression than for the BDI, for which no sound support was found.

The results of both BDI studies were indicative of good convergent validity with other measures of depression. Therefore, one may assume the scale to be an appropriate measure if applied in screening for depression. Nevertheless, high correlations between BDI and measures of disability and anxiety, indicating poor discriminant validity, and the potential impact of the somatic factor identified by Peck et al. (1989) raise concerns. Taking the indeterminate measurement property ratings further into consideration, the meaning of conclusions drawn from the results of both BDI studies is weakened. There is no doubt that the BDI may be a useful tool to assess general feelings of distress, but there is not sufficient evidence for the ability to differentiate between depression and other

psychological problems in RA. It was decided to follow the quality criteria proposed by the COSMIN initiative, relevance and comprehensiveness were only rated positive if all items of a scale referred to the construct of interest or if all aspects of the constructs were covered by the items, respectively. The chosen methodology affects the ratings given for the quality of measurement properties. These harsh and restrictive quality criteria lead to less optimistic results and conclusions concerning the content validity of the measures, as a complete coverage of hundred percent is difficult to attain. If the criteria for relevance and comprehensiveness were loosened, the BDI would have received a positive rating for content validity.

Superior to BDI, the CES-D received positive ratings for structural validity and shortened versions demonstrated adequate criterion validity. Efficiency calculations of Martens et al. (2003; 2006) underline the usability of the full CES-D and shortened versions as a screening tool. According to Martens et al. (2003), a full-scale cutoff-score of 19 was most efficient in identifying cases of MD (sensitivity = .83, specificity = .65). At the expense of specificity, sensitivity increased to .93 with a cutoff-score of 16 (specificity = .42), thus, 19 might not be the most appropriate cutoff-score if one wants to avoid misclassifying patients with MD. In their subsequent study on the development of a shortened CES-D for the identification of MD within RA, Martens et al. (2006) concluded that the shorter version was generally as efficient as the full CES-D. When used with a cutoff-score of 4, a sensitivity value of 1.00 was reported for the shortened scale, meaning that all patients who actually had MD were identified as such. Taken together, the CES-D may be recommended over the BDI for the purpose of screening for depression in RA. Nevertheless, this conclusion is based on a small number of studies and it is not sufficiently assessed to which extent the measures actually differentiate between general psychological distress and depressive symptomatology. As McQuillan et al. (2003) revealed, the CES-D does not differentiate sufficiently between depression and anxiety. Also, the impact of somatic items has to be further illuminated for this specific population as well as optimal cutoff-scores. Future validation studies of depression measures in RA may help to further clarify their utility for various purposes.

Implications for use as outcome measure

Another application is the use as outcome measure for depression. For this purpose, the temporal stability of the measurement structure is relevant to assess change over time. Again, the information gathered on the BDI is not sufficient to recommend it confidently as an accurate measure for this purpose.

In both BDI studies, the temporal stability of the measurement structure and the scales ability and accuracy to detect changes on the construct over time was not assessed. BDI correlations with anxiety measures (STAI and AIMS anxiety scale) were too high to demonstrate adequate discriminant validity and limit the scales' ability to be used as a distinctive measure of depression in RA, it may rather assess a general feeling of distress (Hagglund et al., 1989). Another important finding to take into consideration in the decision to apply the BDI as an outcome measure in RA is the possible contamination through the inclusion of somatic content in the items as also concluded by Peck et al. (1989) from the results of factor analyses and rheumatologists' content rating of items. These items concern appetite, weight loss, not enjoy doing things, effort to do things, sleep difficulties, tiredness, health concerns, and decreased interest in sex. The somatic BDI factor identified in this study may lead to misleading results if interpreted as a measure of depression as scores may be caused through RA disease severity.

Congruent to the recommendations concerning the scales use as a screening measure, there is more evidence for the usability of the CES-D as an outcome measure. Sheehan et al. (1995) reported a stable measurement structure of the CES-D over time and this structure stability was replicated by Rhee et al. (1999). Nevertheless, both studies pointed out that somatic items included in the scale may distort the results and conclusions drawn from scores. Arguing against contamination through somatic items, Martens et al. (2003) found the original CES-D with a cutoff score of 19 to be more efficient than any other cutoff scores of the shortened version. This possible criterion contamination further raises questions whether a single summary score or separate factor scores are more informative.

Implications for future research

The findings of this review highlight shortcomings in research and scientific knowledge about validated PROs for depression in RA. It is remarkable to note that only two commonly used depression measures are validated in their original language for RA. Thus, future research should investigate whether other self-reported depression measures, e.g. the PHQ-9 or the depression subscale of the AIMS, are valid measures of depression in RA. For example, contrary to the BDI and the CES-D, each of the nine PHQ-9 items refers to one of nine DSM-5 criteria for MD, indicating superior content validity of the PHQ-9.

For the BDI, psychometric information was related to construct and structural validity only. The other measurement properties proposed by the COSMIN initiative are still to be examined and validated in RA populations. Although reporting evidence of convergent and divergent validity of the BDI, Peck et al. (1989) also found that this PRO contains two factors, dysphoric mood and somatic complaints, and reflects some somatic contamination in RA populations. Consequently, the DM factor may be a more valid measure of depression in RA, as the SC factor is likely to produce misleading results and may cause overestimations when using a total BDI score. Thus, the potential impact of somatic BDI items needs further clarification; studies designed to assess the criterion validity of a shortened BDI version, excluding somatic items, should clarify whether a modified BDI would result in superior validity. Further, items concerning anxiety may limit the ability to differentiate between separate problems with depression and anxiety. The lack of a positive rating for BDI content validity underlines the need to examine the added value of each separate item.

The same questions still have to be answered for the CES-D. This PRO also contains items that refer to somatic symptoms which may rather be inherently caused and influenced through the rheumatic disease itself. The efforts to develop a shortened version of the CES-D (Martens et al., 2003; Martens et al., 2006; McQuillan et al., 2005) already seem promising and need further confirmation. It is also still unclear whether the CES-D total score is superior to separate factor scores in various situations.

Strengths & limitations

This review, as well as the studies assessed, has its strengths and limitations which should be taken into consideration in future research to enhance the overall quality of validation studies and scientific reviews. On the one hand, use of the COSMIN checklist and corresponding quality criteria proposed by Terwee et al. (2007) is an advantage, as it allows the analysis to be systematically replicated by other researchers. Nevertheless, the use of predefined strict criteria also has disadvantages; studies applying procedures and methods others than those demanded by such predefined criteria may be wrongly deemed to have insufficient quality.

Further limitations may arise from study selection. First of all, validation studies in languages other than the original language version were excluded; their inclusion might have added information and led to different or more precise conclusions concerning the quality of measurement properties of various self-report measures for depression. In this case, the exclusion of studies in other than the original language also led to the exclusion of studies on measures other than the BDI or CES-D. For example, a validation study of the PHQ-9 (Hyphanthis, 2011) was excluded because it assessed the Greek version. Moreover, one cannot rule out the possibility that further studies might have been found and eventually included from databases other than PubMed and Scopus.

During analysis of the included studies, it stood out that the language usage of constructs concerning validity and reliability, as well as statistics in general, varies. Often, required information concerning the applied methods, statistics, or results of a study is not available from the articles and it remains unclear whether things were not done or not reported adequately. These shortcomings in study reports may ultimately distort the results and conclusions drawn from reviews. Standards for the documentation of validation studies may improve the informative quality of research articles and should be applied in future validation studies.

Conclusion

Turning back to the objectives of this review, one may clearly conclude that the extent to which depression measures are validated for use in RA population is scarce. Evidence was only available for the BDI and the CES-D and the quality ratings assigned for the measurement properties result in the conclusion that scientifically sound studies point to the usability of the CES-D rather than the BDI. The area of depression assessment in RA needs further clarification.

Findings for the BDI are based on two studies which both assessed construct and structural validity (Hagglund et al., 1989; Peck et al., 1989). Summed up, the results from correlational and factorial analyses in these two studies indicate adequate convergent validity but poor discriminant validity. Although being a time efficient measure that is simple to administer and score as well as proving adequate psychometric properties in various populations, the BDI also has it caveats. High correlations between the depression and anxiety factors were found, limiting the ability and utility of the BDI to effectively distinguish among separate problems with depression and anxiety in RA in clinical and research settings. Nevertheless, these findings should be interpreted cautiously as measurement properties were rated as indeterminate due to inadequate methodological quality, i.e., both studies failed to formulate precise hypotheses concerning the direction and magnitude of expected correlations with other measures. Also, sample sizes were too small to report sound evidence.

The CES-D received more attention in validation studies so far and taken the measurement property quality ratings into consideration, this PRO seems to be a more valid and reliable measure of depression in RA than the BDI. Nevertheless, this scale also has its difficulties in this specific population where physical symptomatology is common. Despite a stable measurement structure, CES-D scores need to be interpreted with caution due to the possible impact of the somatic factor on total score. To get more differentiated results, researchers and clinicians are advised to calculate separate factor scores to prevent wrong conclusions based on the symptomatology that may wrongly be deemed to be caused by depression.

References

Alamanos, Y., Voulgari, P. V., & Drosos, A. (2006). Incidence and prevalence of rheumatoid arthritis based on the 1987 American College of Rheumatology criteria: A systematic review. *Seminars in Arthritis and Rheumatism*, *36*(3), 182–8. doi:10.1016/j.semarthrit.2006.08.006

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR.* Washington, DC: American Psychiatric Publishing.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Ang, D. C., Choi, H., Kroenke, K., & Wolfe, F. (2005). Comorbid depression is an independent risk factor for mortality in patients with rheumatoid arthritis. *Journal of Rheumatology*, *32*(6), 1013–9. http://www.ncbi.nlm.nih.gov/pubmed/15940760

Angst, J., Gamma, A., Gastpar, M Lepine, J.P., Mendlewicz, J. & Tylee, A. (2002). Gender differences in depression. Epidemiological finding from the European DEPRES I and II studies. European Archives of Psychiatry and Clinical Neurosciences, 252, 201-09. doi:10.1007/s00406-002-0381-6

Astin, J. A., Beckner, W., Soeken, K., Hochberg, M. C., & Berman, B. (2002). Psychological interventions for rheumatoid arthritis: A meta-analysis of randomized controlled trials. *Arthritis and Rheumatism*, *47*(3), 291–302. doi:10.1002/art.10416

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561- 71. doi:10.1001/archpsyc.1961.01710120031004

Beck, A.T., Rush, A.J., Shaw, B.F., Emery, G. (1979). *Cognitive therapy of depression*. New York, NY: Guilford Press.

Beck, A.T., Steer, R.A., Brown, G.K. (1996). *Beck Depression Inventory: Second edition manual*. San Antonio (TX): The Psychological Corporation.

De Graaf, R., ten Have, M. & van Dorsselaer, S. (2010). De psychische gezondheid van de Nederlandse bevolking. NEMESIS-2: Opzet en eerste resultaten. http://www.trefpunteuropa.nl/9353000/1/j4nvgs5kjg27kof_j9vvhswlo9bh5xe/viduev5ba2z4/f=/blg3274.pdf

Dickens, C., & Creed, F. (2001). The burden of depression in patients with rheumatoid arthritis. *Rheumatology, 40*, 1327–30. doi:10.1093/rheumatology/40.12.1327

Dickens, C, McGowan, L, Clark-Carter, D, & Creed, F. (2002). Depression in rheumatoid arthritis: A systematic review of the literature with meta-analysis. *Psychosomatic Medicine, 64*, 52–60. http://www.psychosomaticmedicine.org/content/64/1/52.short

Dominick, K.L., Ahern, F.M., Gold, C.H., & Heller, D.A. (2004). Health-related quality of life among older adults with arthritis. *Health and Quality of Life Outcomes, 2*(5), http://doi.org/10.1186/1477-7525-2-5

Hagglund, K.J., Roth, D.L., Haley, W.E., & Alarcón, G.S. (1989). Discriminant and convergent validity of self-report measures of affective distress in patients with rheumatoid arthritis. *Journal of Rheumatology, 16*(11), 1428-32. http://europepmc.org/abstract/MED/2600941

Hewlett, S., Nicklin, J., Bode, C., Carmona, L., Dures, E., Engelbrecht, M., Hagel, S., Kirwan, J., Molto, A., Redondo, M., & Gossec, L. (2016). Translating patient reported outcome measures: methodological issues explored using cognitive interviewing with three rheumatoid arthritis measures in six European languages. *Rheumatology, 55*, 1009-1016. doi:10.1093/rheumatology/kew011

Kaplan, R.M. & Saccuzzo, D.P. (2009). *Psychological testing. Principles, applications, and issues*(7theds). Belmont, CA: Wadswort.

Kosinski, M., Kujawski, S.C., Martin, R., Wanke, L.A., Buatti, M.C., Ware, J.E., & Perfetto, E.M. (2002). Health-related quality of life in early rheumatoid arthritis: Impact of disease and treatment response. *American Journal of Managed Care, 8*. 231-240. https://www.researchgate.net/publication/11448593/

Kuehner, C. (2003). Gender differences in unipolar depression: An update of epidemiological findings and possible explanations. *Acta Psychiatrica Scandinavica, 108*(3), 163–74. doi: 10.1034/j.1600-0447.2003.00204.x

Löwe, B., Willand, L., Eich, W., Zipfel, S., Ho, A. D., Herzog, W., & Fiehn, C. (2004). Psychiatric comorbidity and work disability in patients with inflammatory rheumatic diseases. *Psychosomatic Medicine, 66,* 395–402. doi:10.1097/01.psy.0000126203.89941.a3

Martens, M. P., Parker, J. C., Smarr, K. L., Hewett, J. E., Ge, B., Slaughter, J. R., & Walker, S. E. (2006). Development of a Shortened Center for Epidemiological Studies Depression Scale for Assessment of Depression in Rheumatoid Arthritis. *Rehabilitation Psychology*, *51*(2), 135–139. doi:10.1037/0090-5550.51.2.135

Martens, M. P., Parker, J. C., Smarr, K. L., Hewett, J. E., Slaughter, J. R., & Walker, S. E. (2003). Assessment of depression in rheumatoid arthritis: a modified version of the center for epidemiologic studies depression scale. *Arthritis and Rheumatism, 49*(4), 549–55. doi:10.1002/art.11203

Martens, M. P., Vandyke, M., Parker, J. C., Smarr, K. L., Hewett, J. E., Hewett, J. E., & Walker, S. E. (2005). Analyzing reliability of change in depression among persons with rheumatoid arthritis. *Arthritis and Rheumatism, 53*(6), 973–8. doi:10.1002/art.21578

McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires* (3rd eds). New York, NY: Oxford University Press.

McQuillan, J., Fifield, J., Sheehan, T. J., Reisine, S., Tennen, H., Hesselbrock, V., & Rothfield, N. (2003). A comparison of self-reports of distress and affective disorder diagnoses in rheumatoid arthritis: a receiver operator characteristic analysis. *Arthritis and Rheumatism, 49*(3), 368–76. doi:10.1002/art.11116

Mokkink, L.B., Terwee, C.B., Knol, D.L., Stratford, P.W., Alonso, J., Patrick, D.L., Bouter, L.M., & de Vet, H.C. (2010a). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Medical Research Methodology, 10*(22). doi:10.1186/1471-2288-10-22

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010b). 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), 539–49. doi:10.1007/s11136-010-9606-8

Morris, A., Yelin, E. H., Panopalis, P., Julian, L., & Katz, P. P. (2011). Long-term patterns of depression and associations with health and function in a panel study of rheumatoid arthritis. *Journal of Health Psychology*, *16*(4), 667–77. doi:10.1177/1359105310386635

Nagyova, I., Stewart, R. E., Macejova, Z., Van Dijk, J. P., & Van den Heuvel, W. J. A. (2005). The impact of pain on psychological well-being in rheumatoid arthritis: The mediating effects of self-esteem and adjustment to disease. *Patient Education and Counseling*, *58*(1), 55–62. doi:10.1016/j.pec.2004.06.011

Newman, S., & Mulligan, K. (2000). The psychology of rheumatic diseases. *Baillière's Clinical Rheumatology*, *14*(4), 773–86. doi:10.1053/berh.2000.0112

Nolen-Hoeksema, S. (1990). *Sex differences in depression*. Stanford, CA: Stanford University Press.

Peck, J.R., Smith, T.W., Ward, J.R., & Milano, R. (1989). Disability and depression in rheumatoid arthritis. A multi-trait, multi-method investigation. *Arthritis & Rheumatism*, *32*(9), 1100–06. http://onlinelibrary.wiley.com/doi/10.1002/anr.1780320908/abstract

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401. https://doi.org/10.1177/014662167001003 06

Rhee, S. H., Petroski, G. F., Parker, J. C., Smarr, K. L., Wright, G. E., Multon, K. D., Buchholz, J.L., & Komatireddy, G. R. (1999). A confirmatory factor analysis of the Center for Epidemiologic Studies Depression Scale in rheumatoid arthritis patients: additional evidence for a four-factor model. *Arthritis Care and Research, 12*(6),392-400. http://www.ncbi.nlm.nih.gov/pubmed/11081010

Santor, D.A. & Coyne, J.C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. Psychological Assessment,9:233–43. doi:10.1037//1040-3590.9.3.233

Sheehan, T.J., & Fifield, J., Reisine, S., & Tennen, H. (1995). The Measurement Structure of the Center for Epidemiologic Studies Depression scale. *Journal of Personality Assessment, 64*(3), 507-21. http://www.tandfonline.com/doi/abs/10.1207/s15327752jpa6403_9

Sheehy, C., Murphy, E., & Barry, M. (2006). Depression in rheumatoid arthritis - Underscoring the problem. *Rheumatology, 45*(11), 1325–7. doi:10.1093/rheumatology/kel231

Sonnenberg, C. M., Beekman, A. T. F., Deeg, D. J. H., & van Tilburg, W. (2000).Sex differences in late-life depression. *Acta Psychiatrica Scandinavica, 101*(4), 286-92. doi: 10.1034/j.1600-0447.2000.101004286.x

Spitzer, R.L., Kroenke, K., Williams, J.B., and the Patient Health Questionnaire Primary Care Study Group (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ (Patient Health Questionnaire) primary care study. *Journal of the American Medical Association, 282*, 1737–44. doi:10.1001/jama.282.18.1737.

Tehlirian, C.V. & Bathon, J.M. (2008). Rheumatoid arthritis. In Klippel, J.H., Stone, J.H., Crofford, L.J., White, P.H (eds). *Primer on Rheumatic Diseases* (13th ed). New York, NY: Springer.

Terwee, C. B., Bot, S. D. M., De Boer, M. R., Van der Windt, D. a W. M., Knol, D. L., Dekker, J., Bouter, L. M., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34–42. doi:10.1016/j.jclinepi.2006.03.012

Terwee, C. B., Jansma, E. P., Riphagen, I. I., & De Vet, H. C. W. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research*, *18*(8), 1115–23. doi:10.1007/s11136-009-9528-5

Timonen, M., Viilo, K., Hakko, H., Särkioja, T., Ylikukju, M., Meyer-Rochow, V. B., Väisänen, E., &Räsänen, P. (2003). Suicides in persons suffering from rheumatoid arthritis. *Rheumatology*, *42*(2), 287–291. doi:10.1093/rheumatology/keg082

Weissman, M.M., Bland, R.C., Canino, G.J., Faravelli, C., Greenwald, S., Hwu, H.G., Joyce, P.R., Karam, E.G., Lee, C.K., Lellouch, J., Lépine, J.P., Newman, S.C., Rubio-Stipec, M., Wells, E., Wickramaratne, P.J., Wittchen, H.U., & Yeh, E.- K. (1996). Cross-national epidemiology of major depression and bipolar disorder. *Journal of the American Medical Association, 276*,(4), 293–99. doi:10.1001/jama.1996.03540040037030.

Wolfe, F., & Michaud, K. (2009). Predicting depression in rheumatoid arthritis: The signal importance of pain extent and fatigue, and comorbidity. *Arthritis and Rheumatism*, *61*(5), 667–73. doi:10.1002/art.24428

Zauszniewski, J.A. & Bekhet, A.K. (2009). Depressive symptoms in elderly women with chronic conditions: Measurement issues. *Aging& Mental Health*, *13*(1):64–72. doi:10.1080/13607860802154481

Zigmond, A.S. & Snaith, R.P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica, 67*, 361–70. doi: 10.1111/j.1600-0447.1983.tb09716.x

Appendix

1.1. PubMed search string

(depression[MeSH] OR depress*[tiab] OR anxiety[MeSH] OR anxi*[tiab] OR fear[MeSH] OR fear[tiab] OR distress[tiab] or worry[tiab] OR angst[tiab] OR sadness[tiab]) AND (Rheumatoid Arthritis[MeSH] OR rheumatoid arthritis[tiab]) AND (instrumentation[sh] OR methods[sh] OR Validation Studies[pt] OR Comparative Study[pt] OR "psychometrics"[MeSH] OR psychometr*[tiab] OR clinimetr*[tw] OR clinometr*[tw] OR "outcome assessment (health care)"[MeSH] OR outcome assessment[tiab] OR outcome measure*[tw] OR "observer variation"[MeSH] OR observer variation[tiab] OR "Health Status Indicators"[Mesh] OR "reproducibility of results"[MeSH] OR reproducib*[tiab] OR "discriminant analysis"[MeSH] OR reliab*[tiab] OR unreliab*[tiab] OR valid*[tiab] OR coefficient[tiab] OR homogeneity[tiab] OR homogeneous[tiab] OR "internal consistency"[tiab] OR (cronbach*[tiab] AND (alpha[tiab] OR alphas[tiab])) OR (item[tiab] AND (correlation*[tiab] OR selection*[tiab] OR reduction*[tiab])) OR agreement[tiab] OR precision[tiab] OR imprecision[tiab] OR "precise values"[tiab] OR test–retest[tiab] OR (test[tiab] AND retest[tiab]) OR (reliab*[tiab] AND (test[tiab] OR retest[tiab])) OR stability[tiab] OR interrater[tiab] OR inter-rater[tiab] OR intrarater[tiab] OR intra-rater[tiab] OR intertester[tiab] OR inter-tester[tiab] OR intratester[tiab] OR intra-tester[tiab] OR interobserver[tiab] OR inter-observer[tiab] OR intraobserver[tiab] OR intra-observer[tiab] OR intertechnician[tiab] OR inter-technician[tiab] OR intratechnician[tiab] OR intra-technician[tiab] OR interexaminer[tiab] OR inter-examiner[tiab] OR intraexaminer[tiab] OR intra-examiner[tiab] OR interassay[tiab] OR inter-assay[tiab] OR intraassay[tiab] OR intra-assay[tiab] OR interindividual[tiab] OR inter-individual[tiab] OR intraindividual[tiab] OR intra-individual[tiab] OR interparticipant[tiab] OR inter-participant[tiab] OR intraparticipant[tiab] OR intra-participant[tiab] OR kappa[tiab] OR kappa's[tiab] OR kappas[tiab] OR repeatab*[tiab] OR ((replicab*[tiab] OR repeated[tiab]) AND (measure[tiab] OR measures[tiab] OR findings[tiab] OR result[tiab] OR results[tiab] OR test[tiab] OR tests[tiab])) OR generaliza*[tiab] OR generalisa*[tiab] OR concordance[tiab] OR (intraclass[tiab] AND correlation*[tiab]) OR discriminative[tiab] OR "known group"[tiab] OR factor analysis[tiab] OR factor analyses[tiab] OR dimension*[tiab] OR subscale*[tiab] OR (multitrait[tiab] AND scaling[tiab] AND (analysis[tiab] OR analyses[tiab])) OR item discriminant[tiab] OR interscale correlation*[tiab] OR error[tiab] OR errors[tiab] OR "individual variability"[tiab] OR (variability[tiab] AND (analysis[tiab] OR values[tiab])) OR (uncertainty[tiab] AND (measurement[tiab] OR measuring[tiab])) OR "standard error of measurement"[tiab] OR sensitiv*[tiab] OR responsive*[tiab] OR ((minimal[tiab] OR minimally[tiab] OR clinical[tiab] OR clinically[tiab]) AND (important[tiab] OR significant[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR meaningful change[tiab] OR "ceiling effect"[tiab] OR "floor effect"[tiab] OR "Item response model"[tiab] OR IRT[tiab] OR Rasch[tiab] OR "Differential item functioning"[tiab] OR DIF[tiab] OR "computer adaptive testing"[tiab] OR "item bank"[tiab] OR "cross-cultural equivalence"[tiab])

1.2. Scopus search string

((TITLE-ABS-KEY(depress* OR anxi* OR fear OR distress OR worry OR angst OR sadness)) AND (TITLE-ABS-KEY("rheumatoid arthritis")) AND (TITLE-ABS-KEY(instrument* OR method* OR psychometr* OR clinimetr* OR clinometr* OR "outcome assessment" OR "outcome measure*" OR "observer variation" OR "health status indicator*" OR "reproduc* of result*" OR "discriminant analys?s" OR reliab* OR unreliab* OR valid* OR coefficient OR homogen* OR "internal consistency" OR agreement OR precision OR imprecision OR "precise value*" OR "test-retest" OR (test AND retest) OR (reliab* AND (test OR retest)) OR stability OR interrater OR inter-rater OR intrarater OR intra-rater OR intertester OR inter-tester OR intratester OR intra-tester OR interobserver OR inter-observer OR intraobserver OR intra-observer OR intertechnician OR inter-technician OR intratechnician OR intra-technician OR interexaminer OR inter-examiner OR intraexaminer OR intra-examiner OR interassay OR inter-assay OR intraassay OR intra-assay OR interindividual OR inter-individual OR intraindividual OR intra-individual OR interparticipant OR inter-participant OR intraparticipant OR intra-participant OR kappa* OR ((replicab* OR repeat*) AND (measure* OR finding* OR result* OR test*)) OR repeatab* OR "cronbach* alpha*" OR "item correlation" OR "item selection" OR "item reduction" OR generali?a* OR concordance OR "intraclass correlation" OR discriminative OR "known group" OR factor analys?s OR dimension* OR subscale* OR "multitrait scaling analys?s" OR item discriminant OR "interscale correlation*" OR error* OR "individual variability" OR "variability analys?s" OR "variability value*" OR "uncertainty measur*" OR "standard error of measurement" OR sensitiv* OR responsiv* OR ((minimal* OR clinical*) AND (important OR significant OR detectable) AND (change OR difference)) OR (small* AND (real OR detectable) AND (change OR difference)) OR "meaningful change" OR "ceiling effect" OR "floor effect" OR "Item response model" OR irt OR rasch OR "Differential item functioning" OR dif OR "computer adaptive testing" OR "item bank" OR "cross-cultural equivalence" OR "crosscultural equivalence")))

2.1. Quality criteria for measurement properties

	+	?	-
Relevance	100% of items refer to DSM-5 MD criteria	≥ 75% of items refer to DSM-5 MD criteria	< 75% of items refer to DSM-5 MD criteria
Comprehen-siveness	All DSM-5 MD criteria covered by the items	≥ 7 DSM-5 MD criteria covered by the items	< 7 DSM-5 MD criteria covered by the items
Construct validity	Specific hypotheses [1] and ≥ 75% of results confirm hypotheses[2]	Doubtful design/ method (e.g., no hypotheses)	< 75% of hypotheses confirmed *
Structural validity	CFA applied and adequate model fit [3]	Doubtful design/ method or no CFA	Model fit not adequate *
Criterion validity	Convincing arguments for GS and correlation with GS >0.70 or AUC ≥ 0.80[4]	No convincing arguments for GS or doubtful design/ method	Correlation with GS <0.70 *
Internal consistency	FA[5] and alpha calculated for each dimension and alpha = 0.70 – 0.95	Doubtful design/method or no FA	alpha <0.70 or >0.95 *

+ = positive; ? = indeterminate; - = negative; n.a. = not applicable; [1] = formulated a priori, including the direction & magnitude of expected correlations; [2] = in sample sizes N ≥ 100; [3] = sample size N 10 x #items & >100; [4] = AUC values as indication of the discriminative ability of a measure; [5] = sample size N 7 x #items & >100; GS = gold standard; * = despite adequate design/ methods.

3.1. Descriptive information of included studies

Author[1]	Title	Year	Source	PRO[2]
Hagglund	Discriminant and convergent validity of self-report measures of affective distress in patients with rheumatoid arthritis.	1989	Journal of Rheumatology	BDI
Martens	Assessment of depression in rheumatoid arthritis: a modified version of the center for epidemiologic studies depression scale.	2003	Arthritis & Rheumatism	CES-D[3]
Martens	Analyzing reliability of change in depression among persons with rheumatoid arthritis.	2005	Arthritis & Rheumatism	CES-D
Martens	Development of a Shortened Center for Epidemiological Studies Depression Scale for Assessment of Depression in Rheumatoid Arthritis.	2006	Rehabilitation Psychology	CES-D[3]
McQuillan	A comparison of self-reports of distress and affective disorder diagnoses in rheumatoid arthritis: a receiver operator characteristic analysis.	2003	Arthritis & Rheumatism	CES-D[3]
Peck	Disability and depression in rheumatoid arthritis. A multi-trait, multi-method investigation.	1989	Arthritis & Rheumatism	BDI
Rhee	A confirmatory factor analysis of the Center for Epidemiologic Studies Depression Scale in rheumatoid arthritis patients: additional evidence for a four-factor model.	1999	Arthritis Care & Research	CES-D
Sheehan	The measurement structure of the Center for Epidemiologic Studies Depression scale.	1995	Journal of Personality Assessment	CES-D

[1] = name of first author, for more information see the reference section; [2] = the depression measure examined in the studies; [3] = modified, shortened versions of the CES-D.

3.2. Measurement properties examined & study population characteristics

Author	PRO	Measurement property	N[1]	% female	Age[2]	Disease duration[2]
Hagglund	BDI	Construct validity, Structural validity	52	61	56,5 (11.9)	13,5
Martens '03	CES-D -SF	Criterion validity	1st phase 160 2nd phase 52	1st phase 64 2nd phase 58	52,8 (21.5) 51,2 (19.8)	13,8 13,6
Martens '05	CES-D	Responsiveness	54	72	54,6 (11.4)	n.r.
Martens '06	CES-D-SF	Criterion validity	337	55	61,0 (12.7)	14,4
McQuillan	CES-D	Construct validity, Criterion validity, Internal consistency	415	83	58,0 (9.7)	10
Peck	BDI	Construct validity, Structural validity	107	63	59,3	17,6
Rhee	CES-D	Structural validity	685 (T1) 537 (T2) 453 (T3)	56,5	59 (12.5)	9[3]
Sheehan	CES-D	Structural validity	988 (T1) 813 (T3)	75	51	n.r.

[1] = all study population consisted of 100% RA patients; [2] = mean years, SD in brackets if reported; [3] = median; n.r. = not reported.

3.3. Linking results of BDI items to DSM-5 criteria for MD

Item nr		Item	DSM-5 MD symptom
1	0	I do not feel sad	Dysphoria
	1	I feel sad	
	2	I am sad all the time and I can't snap out of it	
	3	I am so sad and unhappy that I can't stand it	
2	0	I am not particularly discouraged about the future	- (hopelessness)
	1	I feel discouraged about the future	
	2	I feel I have nothing to look forward to	
	3	I feel the future is hopeless and that things cannot improve.	
3	0	I do not feel like a failure	Worthlessness*
	1	I feel I have failed more than the average person	
	2	As I look back on my life, all I can see is a lot of failures	
	3	I feel I am a complete failure as a person	
4	0	I get as much satisfaction out of things as I used to	Anhedonia
	1	I don't enjoy things the way I used to	
	2	I don't get real satisfaction out of anything anymore	
	3	I am dissatisfied or bored with everything	
5	0	I don't feel particularly guilty	Guilt / Worthlessness
	1	I feel guilty a good part of the time	
	2	I feel quite guilty most of the time	
	3	I feel guilty all of the time	
6	0	I don't feel I am being punished	Guilt / Worthlessness
	1	I feel I may be punished	
	2	I expect to be punished	
	3	I feel I am being punished	
7	0	I don't feel disappointed in myself	Worthlessness
	1	I am disappointed in myself	
	2	I am disgusted with myself	
	3	I hate myself	
8	0	I don't feel I am any worse than anybody else	Worthlessness
	1	I am critical of myself for my weaknesses or mistakes	
	2	I blame myself all the time for my faults	
	3	I blame myself for everything bad that happens	
9	0	I don't have any thoughts of killing myself	Suicidal ideation
	1	I have thoughts of killing myself, but I would not carry them out	
	2	I would like to kill myself	
	3	I would kill myself if I had the chance	
10	0	I don't cry any more than usual	Dysphoria
	1	I cry more now than I used to	
	2	I cry all the time now	
	3	I used to be able to cry, but now I can't cry even though I want to	
11	0	I am no more irritated by things than I ever was	- (irritability)
	1	I am slightly more irritated now than usual	
	2	I am quite annoyed or irritated a good deal of the time	
	3	I feel irritated all the time	
12	0	I have not lost interest in other people	Anhedonia
	1	I am less interested in other people than I used to be	
	2	I have lost most of my interest in other people	
	3	I have lost all of my interest in other people	

13	0 1 2 3	I make decisions about as well as I ever could I put off making decisions more than I used to I have greater difficulty in making decisions more than I used to I can't make decisions at all anymore	Indecisiveness
14	0 1 2 3	I don't feel that I look any worse than I used to I am worried that I am looking old or unattractive I feel there are permanent changes in my appearance that make me look unattractive I believe that I look ugly	Worthlessness
15	0 1 2 3	I can work about as well as before It takes an extra effort to get started at doing something I have to push myself very hard to do anything I can't do any work at all	- (avolition)
16	0 1 2 3	I can sleep as well as usual I don't sleep as well as I used to I wake up 1-2 hours earlier than usual and find it hard to get back to sleep I wake up several hours earlier than I used to and cannot get back to sleep	Insomnia
17	0 1 2 3	I don't get more tired than usual. I get tired more easily than I used to. I get tired from doing almost anything. I am too tired to do anything.	Fatigue/ loss of energy
18	0 1 2 3	My appetite is no worse than usual My appetite is not as good as it used to be My appetite is much worse now I have no appetite at all anymore	Loss of appetite
19	0 1 2 3	I haven't lost much weight, if any, lately I have lost more than five pounds I have lost more than ten pounds I have lost more than fifteen pounds	Loss of appetite/ weight
20	0 1 2 3	I am no more worried about my health than usual I am worried about phys. problems like aches, pains, upset stomach, or constipation I am very worried about physical problems and it's hard to think of much else I am so worried about my physical problems that I cannot think of anything else	- (hypochondria)
21	0 1 2 3	I have not noticed any recent change in my interest in sex I am less interested in sex than I used to be I have almost no interest in sex I have lost interest in sex completely	Anhedonia

*all items related to a negative appraisal of self-worth and/or associated emotions such as increased self-blame/self-hatred categorized as worthlessness.

3.4. Linking results of CES-D items to DSM-V criteria for MD

Item nr	Item	DSM-5 MD symptom
1	I was bothered by things that usually don't bother me.	– (irritability)
2	I did not feel like eating; my appetite was poor.	Loss of appetite
3	I felt that I could not shake off the blues even with help from my family of friends.	Dysphoria
4	I felt I was just as good as other people.	Worthlessness
5	I had trouble keeping my mind on what I was doing.	Concentration difficulties
6	I felt depressed.	Dysphoria
7	I felt that everything I did was an effort.	- (Avolition)
8	I felt hopeful about the future.	- (hopelessness)
9	I thought my life had been a failure.	Worthlessness
10	I felt fearful.	- (anxiety)
11	My sleep was restless.	Insomnia
12	I was happy.	- (dysphoria/reverse)
13	I talked less than usual.	- (social withdrawal)
14	I felt lonely.	- (feeling lonely)
15	People were unfriendly.	- (feeling unloved)
16	I enjoyed life.	- (dysphoria/reverse)
17	I had crying spells.	Dysphoria
18	I felt sad.	Dysphoria
19	I felt that people dislike me.	- (feeling unloved)
20	I could not get "going".	- (avolition)

3.5. Efficiency values of various CES-D versions with different cutoff scores

Scale	Cutoff score	Sensitivity	Specificity	AUC	PPV	NPV
Martens 2003 (items 1,3,5,6,7,11,12,16,18)						
Study 1						
CES-D full scale	16	.93	.42		.55	.88
CES-D full scale	19	.83	.65		.64	.83
CES-D modified	3	.83	.40		.51	.75
CES-D modified	4	.71	.59		.57	.73
CES-D modified	5	.62	.75		.65	.72
CES-D modified	6	.54	.87		.76	.71
CES-D modified	7	.42	.90		.76	.67
CES-D modified	8	.30	.97		.88	.65
Study 2						
CES-D full scale	19	.86	.83		.79	.89
CES-D modified	6	.55	.97		.75	.72
Martens 2006 (items 1,3,5,6,8,10,12,14,18)						
CES-D full scale	16	.91	.87	.95	.52	.98
CES-D full scale	19	.85	.93		.65	.98
CES-D-SF	2	1.00	.52	.94	.25	1.00
CES-D-SF	3	1.00	.63		.30	1.00
CES-D-SF	4	1.00	.73		.37	1.00
CES-D-SF	5	.96	.81		.44	.99
CES-D-SF	6	.83	.87		.50	.97
CES-D-SF	7	.72	.92		.58	.95
McQuillan						
CES-D	16	.89	.24	.92		
CES-D	19	.86	.18			
CES-D (no somatic)	16	.89	.21	*		
CES-D (no somatic)	19	.78	.15			

* = not reported, described in article as “No significant differences between AUC for the full CES-D and the shortened version without somatic items were found”