The validity of self-rating depression scales in patients with chronic widespread pain: a Rasch analysis of the Major Depression Inventory

K Amris1, E Omerovic2, B Danneskiold-Samsøe1, H Bliddal1, EE Wæhrens1,3

1The Parker Institute, Department of Rheumatology, Copenhagen University Hospitals, Bispebjerg and Frederiksberg, Copenhagen, 2Department of Rheumatology, Copenhagen University Hospital, Glostrup, Copenhagen, and 3Research Initiative for Activity Studies and Occupational Therapy, Institute of Public Health, University of Southern Denmark, Odense, Denmark

Background: Assessment of depression in chronic pain patients by self-rating questionnaires developed and validated for use in normal and/or psychiatric populations is common. The aim of this study was to evaluate the psychometric properties of the Major Depression Inventory (MDI) in a sample of females with chronic widespread pain (CWP).

Method: A total of 263 females diagnosed with CWP and referred for rehabilitation completed the MDI as part of the baseline evaluation. Rasch analysis was applied to this dataset. Rasch measurement models allow detailed analyses of an instrument’s rating scale and further aspects of validity, including fit of individual scale items to a unidimensional model indicating assessment of a single construct (depression), as a prerequisite for measurement.

Results: The Rasch analysis revealed substantial problems with the rating scale properties of the MDI and lack of unidimensionality. In contrast to somatic items, MDI items related to depressed mood and negative view of oneself were distributed at the higher end of the item difficulty measurement scale, indicating low endorsement of these items.

Discussion: From the perspective of the Rasch measurement model, the MDI demonstrated insufficient psychometric properties when used to identify and quantify severity of depression in a clinical sample of females with CWP. The observed item endorsement pattern indicated that, in this study population, the relatively high depression severity scores primarily pertained to a common core of pain-related somatic symptoms. Careful consideration when interpreting questionnaire-derived scores of depression implemented in research and routine clinical care of patients with chronic pain is warranted.

Several studies support the association between chronic widespread pain (CWP) and depression (1, 2). In fibromyalgia, the lifetime prevalence of depressive symptoms is reported to be 90%, and 30–86% for major depressive disorder (3), dependent upon criteria for depression (4). The high rate of comorbidity observed between fibromyalgia and major depression has led to consideration of largely overlapping pathophysiologica processes in the central nervous system leading to shared common clinical features (5). When viewed as separate diagnostic entities, they clearly exist in complex bidirectional relationships, such that depression may give rise to altered pain processing and alterations in pain processing may promote affective states conducive to the development of depression (5).

Widely used self-report questionnaires, designed to identify and quantify symptoms of depression, are developed and validated for use in normal and/or psychiatric populations. Built upon the clinical concept of depression, in which somatic symptoms are present but not predominant, their use in chronic pain may be confounded by criterion contamination. Chronic pain patients can acquire a clinically significant score on most depression scales by endorsing items concerning sleep problems, fatigue, and reduced activity – symptoms attributed by patients to pain rather than mood (6, 7). The clinical implications of an uncritical incorporation of standardized depression scales into routine clinical care of chronic pain patients may therefore be an overestimation of depression and misguided intervention. A related issue to the evaluation of depression in pain patients concerns the concept of depression in this context. Typically, patients with clinical depression encountered in psychiatric care display hopelessness, worthlessness, and suicidal thoughts (8). The focus of depression in the context of pain seems to focus more on low mood, somatic symptoms, and the consequences of disability (9, 10). A need for new models describing the complex interaction between depression and chronic pain, and a reconsideration of what constitutes depression in the presence of chronic pain, has been indicated by several authors (6, 9, 11). However, such new models
cannot be derived without the use of valid and reliable assessments of depression in chronic pain populations.

The Major Depression Inventory (MDI) (12) is one of the most commonly used self-rating depression scales in Denmark (13). The MDI is criterion based and provides a potential depression diagnosis according to internationally agreed diagnostic criteria, and a severity score for monitoring the condition. The MDI is recommended by the Danish National Board of Health for routine screening and diagnostic assessment of depression in high-risk categories of primary care patients, including individuals with chronic pain conditions (14). The MDI has also been used in clinical studies on chronic pain populations (15, 16). The psychometric properties of the MDI have been evaluated in mental health (17–19) as well as population-based samples (20), but not in chronic pain populations. The aim of this study was to evaluate the psychometric properties of the MDI when used to identify and quantify the severity of depression among females with CWP.

Method
Sample and data collection

Data were collected from female patients diagnosed with CWP and referred for rehabilitation at the Department of Rheumatology, Frederiksberg Hospital. Ahead of enrolment in the rehabilitation programme, a comprehensive baseline assessment was performed on all patients, including assessment with the MDI and several other self-report and observation-based assessment instruments. Data were collected from 1 March 2007 to 28 February 2009 and stored in a clinical database. All examinations were approved by the local ethics committee (KF 01-045/03). The referral diagnosis of CWP was based on the 1990 American College of Rheumatology (ACR) definition of widespread pain (i.e. pain axially and in a minimum of three body quadrants) (21). Exclusion criteria for the rehabilitation programme were severe physical impairment necessitating assistance in personal activities of daily living (ADL), concurrent history of major psychiatric disorder not related to the pain disorder, and other medical conditions capable of causing patients’ symptoms (e.g. uncontrolled inflammatory/autoimmune disorder, uncontrolled endocrine disorder, and malignancy).

Instruments

The MDI was originally developed in Denmark (12), but has been translated into several languages, including English (see Appendix 1). The MDI was constructed to cover both the ICD-10 and DSM-IV symptoms of depression. A fifth edition of the DSM (DSM-5) was published in May 2013, but in general the criteria for major depressive disorder are identical in DSM-IV and DSM-5 (22). The MDI contains 10 items. Item 8 and 10 are divided into two subitems, a and b, resulting in a total number of 12 items. Each item is scored on a six-category Likert scale according to how much of the time the individual symptom has been present during the past 14 days: 0 representing ‘the symptom has not been present at all’ and 5 representing ‘the symptom has been present all of the time’. When calculating an overall score, only the highest score on items 8a/8b and 10a/10b is used. As a diagnostic instrument, MDI items are dichotomized to indicate the presence or absence of each of the symptoms. In DSM-IV/DSM-5 and ICD-10, the items of depressed mood and lack of interest in daily activities (items 1 and 2) are considered core symptoms of depression. In ICD-10, the lack of energy (item 3) is also considered a core symptom. For diagnostic purposes, items 1–3 are considered significantly present at scores of 4 and 5 (i.e. most of the time, all of the time). For the remaining items (items 4–10), the symptom is considered significantly present at scores of 3–5 (i.e. more than half of the time, most of the time, and all of the time).

The algorithm for DSM-IV/DSM-5 is: items 4 and 5 are combined and only the highest score is considered. Thus, the number of items is nine. Major depression is defined as the presence of at least five of the nine items. However, either item 1 or item 2 should be among the five items. The algorithm for ICD-10 moderate to severe (major) depression is the presence of at least two of the three core symptoms (items 1–3) and at least four of the other seven items. As an assessment instrument, the 10 items are summed with a theoretical score ranging from 0 to 50. Suggested cut-points for mild, moderate, and severe depression in a psychiatric setting are: ≤20, no depression; 21–25, mild depression; 26–30, moderate depression; and ≥31, severe depression (17–19).

Adequate sensitivity and specificity of the MDI algorithm for depression, using clinician-assessed diagnosis of depression as the validity index, are reported in clinical samples of depressed patients (17), but have been found to be low in population-based samples (20). MDI cut-points rather than the diagnostic algorithm have therefore been proposed for population-based studies (20). The internal validity of the MDI has further been evaluated in a smaller sample of depressed patients, based on item response theory (Rasch analysis, Mokken analysis), as well as classical psychometric testing [principal component analysis (PCA)] (19). The study supported acceptable unidimensionality of the MDI scale, but with suboptimal fitting of two somatic items (sleep and appetite) in the Rasch analysis and the lowest factor loading of these items in the PCA (19). No studies evaluating the psychometric properties of the MDI in chronic pain populations are available.

The baseline assessments further included the 36-item Short-Form Health Survey (SF-36), the Fibromyalgia Impact Questionnaire (FIQ), the Generalized Anxiety Disorder Inventory (GAD-10), the catastrophizing
Data analyses

Data analyses were performed using a combined approach. First, Rasch measurement methods were used to analyse the rating scale structure and aspects of validity and reliability including examination of unidimensionality, that is the fit of the MDI items to a unidimensional model indicating assessment of a single construct (depression). Second, to further explore the construct assessed by the MDI in a chronic pain population, construct validity was examined by correlating the MDI sum scores to widely used measures of disease impact in this patient population.

Descriptive statistics and correlation analyses. Data were analysed with SPSS 16.0 for general descriptive statistics and are presented as mean/median, standard deviation (sd)/range, and number of persons in the study population. The Spearman rank order test was used to assess for correlations between the MDI sum score and scores of mental functioning on the FIQ, SF-36, and GAD-10, scores of catastrophic thinking on the CSQ, scores of pain intensity on the FIQ, scores of social functioning and general health on the SF-36, scores of functional ability on the FIQ and SF-36 and observation-based measures of functional ability (AMPS ADL motor and ADL process ability measures), and global scores on the SF-36 and FIQ. We hypothesized that MDI sum scores would be highly correlated (≥ 0.7) to other measures of depression, indicating convergent construct validity, but less correlated (≤ 0.7) to measures of other constructs, indicating divergent construct validity.

Rasch analysis. The following questions were addressed using Rasch analysis:

1. Does the six-category rating scale used with the items of the MDI demonstrate sound psychometric properties?
2. Do the items of the MDI define a single unidimensional construct, namely depression?
3. Do the items of the MDI show a logical hierarchical order of difficulty?
4. Does the MDI provide precise and reliable estimates of depression?

The Partial Credit Model (PCM) (23) and the Rating Scale Model (RSM) (24) are Rasch models used with polytomous data (i.e. data derived from response scales with more than two categories). The models differ only in their assumptions regarding the distance between response categories. While the PCM assumes that the distance between the response categories is not the same for all items, the RSM assumes the distance is the same. The PCM was applied for the MDI, as the likelihood ratio test indicated lack of fit to an interval model (p < 0.001). The model includes two facets (items and persons) and is based on two assertions: (a) the more depressed a person, the more likely that person is to receive higher ratings on more difficult items than is a less depressed person; and (b) the easier an item, the more likely any person is to receive higher ratings than on more difficult items (25). When data meet these expectations, the items and the persons fit the measurement model, supporting internal scale and person response validity, respectively.

The Rasch computer programme WINSTEPS version 3.68.2 (26) was used to implement logarithmic conversions of the ordinal scores of the MDI into equal interval measures of the person’s overall depression severity. Because the conversions are based on log-odds probabilities, the measures of depression severity and the item difficulty measures are expressed in logits (log-odds probability units) (27). WINSTEPS was also used to generate statistics to evaluate aspects of validity and reliability, including fit of the data to the Rasch model assertions (25, 28). Rasch analysis procedures have been described elsewhere in detail (25, 28, 29).

Prior to examining other forms of validity, the performance of the MDI rating scale was evaluated to ensure that it demonstrated sound psychometric properties based on Linacre’s guidelines (30–32). Thus, the frequency distributions across categories should be either uniform or peak in central or extreme categories to signal optimal category use; average category measures should advance monotonically up the rating scale, indicating that persons who are more depressed have higher item ratings (32); scale category outfit mean square (MnSq) values should be ≤ 2.0. Finally, threshold calibrations should also advance monotonically, with no threshold disordering, and thresholds should increase by at least 1.4 logits to show distinction between categories but by no more than 5 logits to avoid large gaps in the variable (31, 32).

To determine whether the MDI items defined a single unidimensional construct (i.e. depression), a PCA of the standardized residuals was performed followed by examination of item goodness-of-fit statistics (30, 33). The PCA of the standardized residuals (i.e. the difference between what the Rasch model predicts and what was observed) was performed to identify possible secondary dimensions within the data. As a rule of thumb, the value of the first contrast should not exceed the size of an eigenvalue expected by chance, usually less than 2 (34, 35).

When analysing goodness of fit, both underfit and overfit to the model was considered (25). Whereas underfit degrades the quality of measures, overfit in general has no practical implications but might be an indication of lack of local independence (i.e. significant
correlations among the items after the contribution of the underlying construct is removed). Furthermore, when assessing an item’s fit to the Rasch model, both infit and outfit statistics were considered. While the infit statistic gives relatively more weight to the performance of persons closer to the item value, the outfit statistic is not weighted and therefore remains more sensitive to the influence of outlying scores. Critical values for mean squares were calculated based on our sample size (36). Items with infit MnSq values > 1.12 or outfit MnSq values > 1.37, combined with z values ≥ 2.0, were considered to underfit (i.e. misfit) and were removed one at a time, in the order of highest MnSq values, considering high infit MnSq values first, as infit underfit is a greater threat to measurement (29). Removal of underfitting items was planned to stop when all items met the criteria for acceptable goodness of fit. Items with infit or outfit MnSq values < 0.6, combined with z values ≥ 2.0, were considered to overfit (25). While such items are not considered a threat to the measurement system, they would be identified but retained in the instrument. Additionally, we planned to investigate whether retaining misfitting items would disrupt the measurement system, by evaluating for differential test functioning (DTF) (37). DTF occurs when measures vary between two versions of a test. The evaluation of DTF related to inclusion or omission of misfitting items was performed by plotting (a) depression severity measures based on a version containing only items with acceptable goodness of fit to the Rasch model and (b) measures based on a version containing all items.

Furthermore, we explored whether the hierarchical order of item difficulties was logical. This was done by comparing (a) the hierarchical order of the MDI item difficulty estimates along the linear scale based on the Rasch analysis to (b) the rank order of the MDI items suggested by Olsen et al based on mean score values for the individual items obtained in a sample of patients with different states of depression (19).

Finally, we evaluated how well the MDI items were targeted to the depression severity level of the sample by examining the item-person map, a graphic display of the distribution of items and the spread of the person’s depression severity measures, generated by the WINSTEPS programme. To evaluate precision and reproducibility of the item difficulty estimates and the depression severity measures, we examined the overall separation and reliability indices. Reliability coefficients have a ceiling of 1.0 but separation coefficients and indices have no ceiling. The separation indices should be at least 2.0 to obtain a desired reliability coefficient of 0.80 for replicability of person and item ordering (25), and the closer the reliability index is to 1.0 (range 0.0–1.0), the better (38). The person separation index was used to calculate the number of distinct levels of depression (strata) that the items could distinguish [strata = (4 × person separation index + 1)/3] (38).

### Results

#### Participants

A total of 271 females diagnosed with CWP entered the rehabilitation programme in the study period and were baseline assessed. Eight had missing data on the MDI and were excluded, leaving a study sample of 263. The median age was 46 years (range 20.4–71.5) and median symptom duration 85 months (range 6–540). In addition to CWP, 257 (97.7%) of the participants had a tender point count of ≥ 11, thereby also fulfilling the dual 1990 ACR classification criteria for fibromyalgia (21).

#### Sample characteristics and MDI scores

Sample characteristics and MDI classification according to the MDI are summarized in Table 1. The mean total score

| Age (years)          | 45.4 (9.7) | 46.0 (20.4–71.5) |
|----------------------|------------|------------------|
| Duration of pain (months) | 120.64 (101.9); 85.0 (6–540) |
| Living alone         | 88 (38)    |
| Employed             | 72 (27)    |
| Level of education: high school or higher | 100 (43) |
| Tender point count   | 16.7 (2.3); 18.0 (5–18) |
| FIQ VAS pain         | 7.0 (2.0); 7.5 (1–10) |
| FIQ function         | 5.3 (2.2); 5.7 (0–9) |
| SF-36 physical function | 43.0 (20.2); 40.0 (0–95) |
| AMPS motor           | 1.07 (0.50); 1.06 (0.04–2.82) |
| AMPS process         | 1.09 (0.35); 1.09 (0.12–2.18) |
| SF-36 social functioning | 48.2 (26.5); 50.0 (0–100) |
| SF-36 general health  | 31.5 (18.3); 30.0 (0–97) |
| SF-36 MCS            | 40.7 (12.0); 41.0 (14.9–66.6) |
| SF-36 PCS            | 27.0 (6.7); 27.1 (8.8–50.7) |
| FIQ total            | 61.4 (18.7); 62.0 (2.9–97.4) |
| GAD-10 anxiety score | 19.2 (9.8); 18.0 (1–50) |
| CSQ catastrophizing  | 15.9 (8.4); 15.0 (0–36) |
| SF-36 mental well-being | 55.6 (20.6); 56.0 (0–100) |
| FIQ VAS depression   | 3.8 (3.4); 3.0 (0–10) |
| FIQ VAS anxiety      | 4.5 (3.4); 4.4 (0–10) |
| MDI sum score        | 21.6 (10.7); 20.0 (3–50) |

MDI classification based on ICD-10:

| No depression (≤ 20) | 133 (51) |
| Mild depression (21–25) | 44 (17) |
| Moderate depression (26–30) | 27 (10) |
| Severe depression (≥ 31) | 59 (22) |

MDI classification based on ICD-9:

| No depression | 186 (71) |
| Mild depression | 17 (6) |
| Moderate depression | 34 (13) |
| Severe depression | 26 (10) |

MDI, Major Depression Inventory; FIQ, Fibromyalgia Impact Questionnaire; VAS, visual analogue scale; SF-36, 36-item Short-Form Health Survey; AMPS, Assessment of Motor and Process Skills; MCS, mental composite score; PCS, physical composite score; GAD-10, Generalized Anxiety Disorder Questionnaire; CSQ, Coping Strategy Questionnaire; AMPS process, AMPS activities of daily living (ADL) process ability measure; AMPS motor, AMPS ADL motor ability measure. Values are given as n (%) or mean (sd) and median (range).
on the MDI for the sample was 21.6 (sd = 10.7; range 3–50). Applying the conventional cut-points from a psychiatric setting, 49% of the study population could be classified as having clinical depression (mild, moderate, severe), and 32% major depression (mild, moderate, severe). A total of 29% were classified as having clinical depression (mild, moderate, severe) and 23% major depression according to the ICD-10 algorithm. Severe depression was present in 22% of the participants according to the cut-point classification vs. 10% according to the ICD-10 algorithm.

Rasch analyses

*Rating scale analysis.* In the initial analysis, scale category thresholds were disordered in all items. In a subsequent analysis, we therefore collapsed the ‘slightly less than half of the time’ category (score = 2) with the ‘slightly more than half of the time’ category (score = 3) into a new rating scale category ‘half of the time’, and thereby eliminated the threshold disordering in all but two items [items 4 (‘less self-confident’) and 6 (‘felt that life wasn’t worth living’)]. The remaining analyses were based on this five-category rating scale. The frequency distribution typically peaked in categories 2 and 3, the order of category measures was acceptable and, except for items 9 (‘trouble sleeping’), 10a (‘reduced appetite’), and 10b (‘increased appetite’), category outfit MnSq values were < 2.0. None of the items displayed a scale structure of at least 1.4 logits between thresholds, suggesting lack of distinction between categories. In addition, the item-person distribution map revealed gaps in the variable, indicating poorly defined or tested regions of the variable (Figure 1).

| Measure | More depressed persons | More difficult items |
|---------|------------------------|---------------------|
| <more> | Items: Bottom | Items: Center | Items: Top |
| 6 | # | + | + | 6 |
| 5 | + | + | 5 |
| 4 | # | | | 4 |
| 3 | + | + | 3 |
| 2 | # | + | | 2 |
| 1 | # | | | 1 |
| 0 | . # | mean of items | X | 0 |
| -1 | # | + | + | -1 |
| -2 | # | + | | -2 |
| -3 | X | + | X | -3 |
| -4 | X | + | X | -4 |
| -5 | . # | | | -5 |

Less depressed persons

| Measure | Less difficult items |
|---------|---------------------|
| <less> | |

Less difficult items

| Measure | <frequent> |
|---------|------------|

Figure 1. Item-person distribution map based on 12 items and the five-category scale. The most difficult items and the most depressed persons are at the top. Each item is shown in three columns representing different rating scale measures. Items: centre = the mean item difficulty calibration; Items: Bottom = measure level corresponding to the probability of 0.5 of being rated in (or exceeding) the lowest category of the rating scale; Items: Top = measure level corresponding to a probability of 0.5 of being rated in (or falling below) the highest category of the rating scale.
Table 2. Item measurement report for the initial Rasch analysis.

| Item | Count (median score) | Item difficulty measure | Infit | Outfit |
|------|----------------------|-------------------------|-------|--------|
| 6    | 263 (0)              | 0.96                    | 0.08  | 0.71   | -2.8  | 0.50  | -2.7 |
| 10b  | 263 (0)              | 0.92                    | 0.08  | 2.14   | 8.3   | 2.30  | 6.2  |
| 10a  | 263 (1)              | 0.07                    | 1.43  | 4.0    | 1.65  | 4.1   |
| 8b   | 263 (1)              | 0.08                    | 0.78  | -2.8   | 0.76  | 1.7   |
| 4    | 263 (1)              | 0.08                    | 0.76  | -3.1   | 0.76  | -2.8  |
| 8a   | 263 (1)              | 0.08                    | 0.95  | -1.8   | 0.85  | -1.8  |
| 1    | 263 (1)              | 0.07                    | 1.13  | 1.5    | 1.12  | 1.3   |
| 5    | 263 (1)              | 0.09                    | 0.93  | -2.2   | 0.94  | -2.1  |
| 7    | 263 (1)              | 0.07                    | 0.64  | -4.8   | 0.64  | -4.8  |
| 2    | 263 (2)              | 0.07                    | 0.64  | -4.9   | 0.64  | -4.9  |
| 9    | 263 (3)              | -1.08                   | 0.07  | 1.50   | 5.3   | 1.60  | 6.0  |
| 3    | 263 (4)              | -1.27                   | 0.09  | 0.90   | -1.2  | 0.88  | -1.4 |

Infit, Standard error; MnSq, mean square.
Items listed according to item difficulty. Underfitting items appear in bold. Overfitting items appear in italic.

Unidimensionality. In the initial PCA of the standardized residuals, the measures explained 53.1% of the variance, indicating a lack of unidimensionality. After removal of the two items contributing to the largest variance (items 5, 9, 10a, and 10b), the measure explained 45.4% of the variance, indicating improved unidimensionality. After removal of the remaining eight items displayed underfitting, the item difficulties were not contributing to the measurement system, as evidenced by the large standard errors (i.e., imprecision of the measure). The results revealed that the measures based on the reduced eight-item version now explained 65.7% of the variance, indicating improved unidimensionality. After removal of the remaining eight items displayed underfitting, none of the underlying construct. The underfitting items were removed, one at a time. After removal of the two underfitting items related to disturbed appetite, another underfitting item was included, indicating a lack of unidimensionality. Furthermore, item 6, referring to disturbed appetite, displayed a lack of unidimensionality (Table 2).
Hierarchical order of item difficulty. The distribution of the item difficulty measures (i.e. average item calibration) on the logit scale are presented in Table 2. In the current analyses, items with a negative calibration are the easier items, whereas items distributed at the other end of the scale (i.e. with a positive calibration) are the most difficult. The results of the Rasch analysis show that most items related to depressed mood and negative view of oneself were distributed at the higher (positive) end of the item difficulty measurement scale, indicating low endorsement of these items in the study population. Item 6 (‘felt that life wasn’t worth living’) obtained the highest positive calibration score, consistent with a mean score of 0 on that item in the overall study population. Items related to lack of energy and interest in daily activities, sleep disturbance, and difficulties concentrating were all distributed at the lower (negative) end of the item difficulty measurement scale, indicating high endorsement of these items in the study population. Items 3 (‘felt lacking in energy and strength’) and 9 (‘trouble sleeping’) had the lowest negative calibration measures, consistent with an observed median score of 4 and 3, respectively, in the study population. The removal of underfitting items did not affect the hierarchical order of the remaining items (see online Supplementary Table S1).

The hierarchical order of items obtained by Rasch analysis in our study population differed from the rank order of items reported by Olsen et al based on mean score values in a psychiatric sample with a range of depression severity (Table 3). Corresponding to the findings in our CWP population, the most endorsed item (i.e. the highest ranked item) in the depressed sample was item 3 (‘felt lacking in energy and strength’) and the least endorsed item (i.e. the lowest ranked item) was item 6 (‘felt that life wasn’t worth living’). Otherwise, the structure of inclusiveness showed a different pattern in the depressed sample with higher mean scores and ranking of items related to depressed mood and negative view of oneself than in our CWP population. Item 9 (‘trouble sleeping’) was one of the lowest ranked (i.e. least endorsed) items in the depressed sample.

Reliability and precision. As mentioned previously, the item-person distribution map (Figure 1) illustrates that the items and participants were not distributed evenly across the scale. The targeting of the 12 – item version of the MDI to the participants’ depression severity (mean item difficulty estimate: zero, sd = 0.90; mean depression severity estimate: −0.87, sd = 1.89) indicated that the participants had a slightly lower level of depression than the mean item difficulty estimate (zero by default). As the most depressed persons are positioned below the top category of the most difficult item and all persons are above the bottom category of the easiest item, there is no indication of a floor or ceiling effect in this sample. The initial separation index was 2.33 (reliability = 0.84) for persons and 7.85 (reliability = 0.98) for items. Based on the person separation, the number of levels of depression that the MDI could distinguish was 3.4; that is, more than three ranges could be differentiated. After removal of underfitting items, the person separation index increased slightly from 2.33 to 2.68 (reliability = 0.88) and the item separation index increased from 7.85 to 9.28 (reliability = 0.99).

Construct validity

Relationships between MDI scores and other disease variables are summarized in Table 4. The MDI sum score showed a very strong relationship with the FIQ visual analogue scale (VAS) score for depression, the
Validity of the MDI in CWP

GAD-10 sum score for generalized anxiety, the SF-36 score for mental well-being, and the SF-36 mental composite score (MCS), the strength of the correlation ranging from $r = 0.72$ to $0.85$, indicating convergent construct validity. Relationships between the MDI sum score and pain and self-reported physical and social functioning on the FIQ and SF-36 ranged from $r = 0.34$ to $0.54$, indicating divergent construct validity. The MDI sum score showed only a trivial relationship with the observation-based measures of functional ability (AMPS).

Discussion

Although the MDI has demonstrated adequate psychometric properties when applied in mental health and background populations, the data presented here suggest problems with the scaling properties and unidimensionality of the MDI used to identify and quantify severity of depression in female patients with CWP and fibromyalgia.

The MDI can be used both as a diagnostic instrument and as an assessment instrument based on a sum score. A prerequisite for both is that the rating scale demonstrates sound psychometric properties. Applied in our study population, the six-category Likert scale, which forms the basis of data collection, had identifiable problems in the form of threshold disordering. The rating scale categories ‘slightly more than half of the time’ (score = 3) and ‘slightly less than half of the time’ (score = 2) did not apply and had to be collapsed into a new category termed ‘half of the time’ to achieve sufficient rating scale properties. Threshold disordering at this particular level may represent a serious threat to the validity of the diagnostic algorithm, where MDI items are dichotomized to indicate the presence or absence of symptoms, setting the cutoff at a category score of three. Use of the MDI diagnostic algorithm in CWP populations therefore requires careful consideration.

Unidimensionality refers to the existence of one underlying construct that accounts for variation in examinee responses and is a prerequisite for measurement. The result of our study demonstrates the misfit of four items (related to appetite, sleep, and feelings of guilt/bad conscience) suggesting a lack of unidimensionality, that is that the MDI does not measure a single construct (depression) when applied in CWP populations. Evaluation of the severity of depression based on summed scores and classification of depression-based on cut-points should therefore be interpreted with caution or simply avoided in this patient population. Sub-optimal fit of items related to appetite and sleep has also been reported, when the MDI has been subjected to Rasch analysis in samples of depressed patients (19), indicating that these items may represent another construct across different patient populations. Moreover, the results of our first analysis including all 12 items showed that item 6 (‘felt that life wasn’t worth living’) had outfit statistics below 0.6, indicating that this item may not be contributing to the measurement of the underlying construct. As this item is considered a core symptom of depression and therefore included in the diagnostic algorithm, this further undermines the validity of the MDI used as a diagnostic instrument in CWP populations.

The MDI classified 32% of our study population with major (moderate, severe) depression based on conventional cut-points vs. 23% when using the ICD-10

Table 4. Relationships between MDI scores and disease variables.

| Spearman rho correlations | rho   | p     | n   |
|---------------------------|-------|-------|-----|
| MDI sum score: correlation with FIQ depression | 0.72  | 0.000 | 263 |
| MDI sum score: correlation with FIQ anxiety | 0.61  | 0.000 | 262 |
| MDI sum score: correlation with GAD-10 anxiety | 0.85  | 0.000 | 263 |
| MDI sum score: correlation with SF-36 mental wellbeing | 0.77  | 0.000 | 263 |
| MDI sum score: correlation with CSQ catastrophizing | 0.58  | 0.000 | 262 |
| MDI sum score: correlation with FIQ VAS Pain | 0.35  | 0.000 | 263 |
| MDI sum score: correlation with SF-36 social functioning | 0.54  | 0.000 | 263 |
| MDI sum score: correlation with SF-36 general health | 0.42  | 0.000 | 262 |
| MDI sum score: correlation with SF-36 PCS | 0.06  | 0.306 | 263 |
| MDI sum score: correlation with SF-36 MCS | 0.74  | 0.000 | 263 |
| MDI sum score: correlation with FIQ total score | 0.67  | 0.000 | 263 |
| MDI sum score: correlation with FIQ function | 0.44  | 0.000 | 262 |
| MDI sum score: correlation with SF-36 function | 0.34  | 0.000 | 263 |
| MDI sum score: correlation with AMPS motor | 0.17  | 0.006 | 250 |
| MDI sum score: correlation with AMPS process | 0.15  | 0.016 | 250 |

MDI, Major Depression Inventory; FIQ, Fibromyalgia Impact Questionnaire; GAD-10, Generalized Anxiety Disorder Questionnaire; SF-36, 36-item Short-Form Health Survey; CSQ, Coping Strategy Questionnaire; VAS, visual analogue scale; PCS, physical composite score; MCS, mental composite score; AMPS, Assessment of Motor and Process Skills; AMPS process, AMPS activities of daily living (ADL) process ability measure; AMPS motor, AMPS ADL motor ability measure.
algorithm. However, the obtained hierarchical order of item difficulty (i.e. the position of the scale items on a continuum from less difficult to more difficult) showed a characteristic pattern that differed from the symptom endorsement pattern obtained with the MDI in a sample of non-pain patients with different states of depression (19). In our sample, the more difficult items (i.e. the least endorsed items) were mainly items related to depressed mood and negative view of oneself whereas the easier items (i.e. the most endorsed items) were related to lack of energy, poor sleep, difficulty concentrating, and loss of interest in daily activities. Looking closely at the sign of calibration (Table 2), most items related to mood and negative view of oneself had positive calibrations (were the less endorsed) whereas items that could pertain to the pain condition per se had negative calibrations (i.e. were highly endorsed by the study population).

The negative impact of CWP on the ability to perform ADL is well established (39). We could therefore speculate that a relatively high endorsement of item 2 (‘lost interest in daily activities’) in our study population, rather than reflecting lack of interest in daily activities due to depression, pertained to the respondents experience of a substantial pain-related interference with functional ability. Poor sleep, fatigue, and difficulties concentrating are common in chronic pain patients and considered core symptoms in fibromyalgia (40, 41). MDI items covering these symptoms were the most endorsed items in our study population. This symptom endorsement pattern could indicate that the relatively high MDI scores obtained in our study population were primarily due to a common core of pain-related somatic symptoms unrelated to mood. In the depressed population, the structure of inclusiveness showed that the core symptoms of depression were among the most inclusive items whereas (for instance) poor sleep was ranked in the bottom (19).

The notion of a common core of pain-related somatic symptoms contributing to high scores on depression rating scales in chronic pain patients is supported by a study analysing the factor structure of the Beck Depression Inventory (BDI) in a large sample of chronic pain patients (11). In that study, two factors emerged. Factor 1 consistently included items reflecting negative evaluations of the self, on which most patients scored relatively low. Items loading on factor 2 were predominantly concerned with somatic and physical function, on which most patients scored at least moderately. Based on these findings, Morley et al concluded that the factor structure meant that, despite relatively high total BDI scores in the sample, the content of the scores showed relatively little of the cognitive beliefs characterizing depressed patients without chronic pain (11). Instruments omitting somatic and some cognitive items (guilt, suicidal thoughts), such as the widely used Hospital Anxiety and Depression Scale (HADS) (42), have been developed to assess anxiety and depression in individuals with somatic illnesses and are recommended for use in chronic pain populations (8). However, psychometric studies indicate that, although the HADS may be a clinical useful scale of emotional distress, its ability to differentiate between the constructs of anxiety and depression is unclear, which means that its use needs to be targeted to more general assessment of distress (43).

The results of such psychometric studies emphasize the need for careful consideration when interpreting questionnaire-derived scores of depression implemented in the clinic and in research, dedicated to discerning the nature of the relationship between depression and chronic pain. This may in particular apply to patients with fibromyalgia, as the observed symptom endorsement pattern on the MDI showed a substantial overlap with the aggregate of symptoms suggested as the core of the proposed symptom-based diagnostic and survey criteria for this pain condition (40, 44). The operationalized depression domain in the MDI is conceptualized as comprising a set of symptoms covering the ICD-10 and DSM-IV symptoms of depression. However, the observed very strong association between the MDI score and scores of anxiety and general mental wellbeing on other self-report instruments could indicate that, rather than being a score of the domain of depression, the MDI score reflected a more general score of emotional distress when applied in our study population. The observed associations between the MDI score and pain, and self-reported physical and social functioning, were only moderate, explaining about 12–33% of the variance between variables, suggesting assessment of different constructs.

Overall, the MDI showed adequate precision and reproducibility of the item difficulty estimates and depression severity measures, when applied in our study population. The Rasch analysis supported the possibility of developing a revised unidimensional eight-item version of the MDI, based on a five-category Likert scale, but removal of underfitting items did not change the item endorsement pattern. The revised version of the MDI would provide linear severity measures of a unidimensional construct of some sort in CWP populations, presumably pain-related distress. Use of this instrument to identify and quantify the depression severity among individuals with CWP would therefore require further validation.

The study was limited by only including women. It was therefore not possible to include an examination of gender-related differential item functioning (i.e. examination of differences in item difficulty estimates across gender) in the Rasch analysis. A significant gender difference in the prevalence of an MDI score ≥ 20 has been reported in a Danish epidemiological survey, where no significant gender difference in the prevalence of major depression was present (45). Participants were recruited
from tertiary care, and although a wide range in MDI scores was obtained, this may have influenced the outcome of the Rasch analysis and the generalizability of the study results.

In conclusion, from the perspective of the Rasch rating scale model, the 12-item version of the MDI, based on a six-category Likert scale, demonstrated insufficient psychometric properties when applied in a clinical sample of female patients with CWP and fibromyalgia. The Rasch analysis revealed problems with the rating scale properties as well as lack of unidimensionality, suggesting a serious threat to the validity of the inherent diagnostic algorithm and depression severity scores provided by the instrument in this patient population. Moreover, the Rasch analysis demonstrated a characteristic item endorsement pattern, indicating that the relatively high scores on the MDI, rather than pertaining to depressed mood and negative view of oneself, were primarily related to a common core of pain-related somatic items. Thus, use of the MDI to identify and quantify the severity of depression among females with CWP cannot be recommended.

Acknowledgements
We thank the research staff at the Parker Institute, the staff at the Department of Rheumatology, and the occupational therapists at the Department of Occupational Therapy, Frederiksberg Hospital for contributing to the data collection.

This study was supported by grants from the Oak Foundation, Aase og EjnarDanielsen fond, the Schioldann Foundation, the Danish Rheumatism Association, the A.P. Møller Foundation for the Advancement of Medical Science, Minister Erna Hamilton’s Foundation for the Arts and Sciences, and the association supporting treatment and research in rheumatologic diseases at Frederiksberg Hospital.

References
1. Dworkin SF, Von Korff M, LeResche L. Multiple pains and psychiatric disturbance. An epidemiologic investigation. Arch Gen Psychiatry 1990;47:239–44.
2. Gupta A, Silman AJ, Dey D, Morriss R, Dickens C, Macfarlane GJ, et al. The role of psychosocial factors in predicting the onset of chronic widespread pain: results from a prospective population-based study. Rheumatology (Oxford) 2007;46:666–71.
3. Gracely RH, Ceko M, Bushnell MC. Fibromyalgia and depression. Pain Rev Treat 2012;2012:486590.
4. Aaron LA, Bradley LA, Alarcon GS, Alexander RW, Triana-Alexander M, Martin MY, et al. Psychiatric diagnoses in patients with fibromyalgia are related to health-care-seeking behavior rather than to illness. Arthritis Rheum 1996;39:436–45.
5. Maletic V, Raison CL. Neurobiology of depression, fibromyalgia and neuropathic pain. Front Biosci 2009;14:5291–338.
6. Pincus T, Williams A. Models and measurements of depression in chronic pain. J Psychosom Res 1999;47:211–19.
7. Turk DC, Okifuji A. Detecting depression in chronic pain patients: adequacy of self-reports. Behav Res Ther 1994;32:9–16.
8. Rusu AC, Pincus T, Morley S. Depressed pain patients differ from other depressed groups: examination of cognitive content in a sentence completion task. Pain 2012;153:1898–904.
9. Nicholas MK, Coulston CM, Ashgari A, Malhi GS. Depressive symptoms in patients with chronic pain. Med J Aust 2009;190: S66–S70.
10. Pincus T, Morley S. Cognitive-processing bias in chronic pain: a review and integration. Psychol Bull 2001;127:599–617.
11. Morley S, Williams AC, Black S. A confirmatory factor analysis of the Beck Depression Inventory in chronic pain. Pain 2002;99:289–98.
12. Bech P, Stage KB, Nair NP, Larsen JK, Kragh-Sørensen P, The Major GA. Depression Rating Scale (MDS). Inter-rater reliability and validity across different settings in randomized moclobemide trials. Danish university antidepressant group. J Affect Disord 1997;42:39–48.
13. Christensen KS, Sokolowski I, Olesen F. Case-finding and risk group screening for depression in primary care. Scand J Prim Health Care 2011;29:80–4.
14. [Guidelines for treatment of unipolar depression in adults], in Danish. Danish Health and Medicines Authority. (https://sundhedsstyrelsen.dk/da/udgivelser/2007/referenceprogram-for-unipolar-depression-hos-vokse) Accessed 9 September 2015.
15. Bergmann N, Ballegaard S, Bech P, Hjalmarsen A, Krogh J, Gyntelberg F, et al. The effect of daily self-measurement of pressure pain sensitivity followed by acupuncture on depression and quality of life versus treatment as usual in ischemic heart disease: a randomized clinical trial. PLoS One 2014;9:e97553.
16. Gormsen L, Rosenberg R, Bach FW, Jensen TS. Depression, anxiety, health-related quality of life and pain in patients with chronic fibromyalgia and neuropathic pain. Eur J Pain 2010;14:127–8.
17. Bech P, Rasmussen NA, Olsen LR, Ahlberg V, Ahlbgard W. The sensitivity and specificity of the major depression inventory, using the present state examination as the index of diagnostic validity. J Affect Disord 2001;66:159–64.
18. Cuipers P, Dekker J, Noteboom A, Smits N, Peen J. Sensitivity and specificity of the major depression inventory in outpatients. BMC Psychiatry 2007;7:39–44.
19. Olsen LR, Jensen DV, Noerholm V, Martiny K, Bech P. The internal and external validity of the major depression inventory in measuring severity of depressive states. Psychol Med 2003;33:351–6.
20. Forsell Y. The major depression inventory versus schedules for clinical assessment in neuropsychiatry in a population sample. Soc Psychiatry Psychiatr Epidemiol 2005;40:209–13.
21. Wolfe F, Smythe HA, Yunus MB, Bennett R, Bombardier C, Goldenberg DL, et al. The American college of rheumatology 1990 criteria for the classification of fibromyalgia. Report of the multicenter criteria committee. Arthritis Rheum 1990;33:160–72.
22. Uher R, Payne JL, Pavlova B, Perlis RH. Major depressive disorder in DSM-5: implications for clinical practice and research of changes from DSM-IV. Depress Anxiety 2014;31:459–71.
23. Masters G. A rasch model for partial credit scoring. Psychometrika 1982;47:149–74.
24. Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43:561–73.
25. Bond TG, Fox CM. Applying the Rasch model. Fundamental measurement in the human science, 2nd edn. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2007.
26. WINSTEPS® Rasch measurement computer program [computer program]. Version 3.68.2. Chicago, IL, 2009.
27. Wright BD. Logits? Rasch Measurement Transactions 1993;7:288.
28. Wright BD, Masters GN. Rating scale analysis. Rasch measurement, Chicago, IL: MESA Press, 1982.
29. Linacre JM. A User’s Guide to W I N S T E P S® M I N I S T E P Rasch-Model Computer Programs. Program manual. Chicago, IL: winsteps.com; 2006.
30. Linacre JM, Wright BD. Construction of measures from many-facet data. J Appl Meas 2002;3:486–512.
31. Linacre JM. Investigating rating scale category utility. J Outcome Meas 1999;3:103–22.
32. Linacre JM. Category disordering vs. step (threshold) disordering. Rasch Measurement Transactions 1999;13:675.
33. Smith EV Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Applied Measurement 2002;3:205–31.
34. A users guide to Winsteps(R) Ministep Rasch-model computer programs.Program manual 3.7.5.0 [computer program]. 2012.
35. Raîche G. Critical eigenvalue sizes (variance) in standardized residual principal component analysis (PCA). Rasch Measurement Transactions 2005;19:1012.
36. Smith RM, Schumacker RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. J Outcome Meas 1998;2:66–78.
37. Linacre JM. Differential item and test functioning (DIF & DTF). Rasch Measurement Transactions 2002;16:889.
38. Fisher WP. Reliability statistics. Rasch Measurement Transactions 1992;6:238.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Table S1. Item measurement report for the Rasch analysis after removal of underfitting items.

Figure S1. Category structure for item 1: Have you felt low in spirits or sad?

Annex 1. Major Depression Inventory.

Annex 2

Please note that the editors are not responsible for the content or functionality of any supplementary material supplied by the authors. Any queries should be directed to the corresponding author.