



ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

## ARCHIVIO ISTITUZIONALE DELLA RICERCA

### Alma Mater Studiorum Università di Bologna Archivio istituzionale della ricerca

The Hamilton Rating Scales for Depression: A critical review of clinimetric properties of different versions

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Carrozzino D., Patierno C., Fava G.A., Guidi J. (2020). The Hamilton Rating Scales for Depression: A critical review of clinimetric properties of different versions. *PSYCHOTHERAPY AND PSYCHOSOMATICS*, 89(3), 133-150 [10.1159/000506879].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/762175> since: 2024-06-07

*Published:*

DOI: <http://doi.org/10.1159/000506879>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Carrozzino, D., Patierno, C., Fava, G. A., & Guidi, J. (2020). The Hamilton Rating Scales for Depression: A Critical Review of Clinimetric Properties of Different Versions. *Psychotherapy and Psychosomatics*, 89(3), 133–150.

The final published version is available online at: <https://doi.org/10.1159/000506879>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

**When citing, please refer to the published version.**

**The Hamilton Rating Scales for Depression:  
A Critical Review of Clinimetric Properties of Different Versions**

Danilo Carrozzino<sup>a</sup>, Chiara Patierno<sup>a</sup>, Giovanni A. Fava<sup>b</sup>, Jenny Guidi<sup>a</sup>

<sup>a</sup>Department of Psychology, University of Bologna, Bologna, Italy

<sup>b</sup>Department of Psychiatry, University at Buffalo, State University of New York, Buffalo, NY, USA

**Running title:** Hamilton Scales

**Corresponding author:**

Danilo Carrozzino, PhD

Department of Psychology, University of Bologna

Viale Berti Pichat 5,

40127 Bologna (Italy)

E-mail: [danilo.carrozzino@unibo.it](mailto:danilo.carrozzino@unibo.it)

## **Abstract**

The format of the original Hamilton Rating Scale for Depression (HAM-D) was unstructured: only general instructions were provided for rating individual items. Over the years, a number of modified versions of the HAM-D have been proposed. They differ not only in the number of items, but also in modalities of administration. Structured versions, including item definitions, anchor points and semi-structured or structured interview questions, were developed. This comprehensive review was conducted to examine the clinimetric properties of the different versions of the HAM-D. The aim was to identify the HAM-D versions that best display the clinimetric properties of reliability, validity, and sensitivity to change. The search was conducted on MEDLINE, Scopus, Web of Science, and PubMed, and yielded a total of 35,473 citations, but only the most representative studies were included. The structured versions of the HAM-D were found to display the highest inter-rater and test-retest reliability. The Clinical Interview for Depression (CID) and the 6-item HAM-D (HAM-D<sub>6</sub>) showed the highest sensitivity in differentiating active treatment from placebo. The findings indicate that the HAM-D is a valid and sensitive clinimetric index, which should not be discarded in view of obsolete and not clinically relevant psychometric criteria. The HAM-D, however, requires an informed use: unstructured forms should be avoided and the type of HAM-D version that is selected should be specified in the registration of the study protocol and in the methods of the trial.

**Keywords:** Clinimetrics; Depression; Hamilton Rating Scales; Clinical Interview for Depression; Symptom Questionnaire; Clinical Pharmacopsychology; Antidepressant Drugs; Psychotherapy; Meta-analysis; Randomized Controlled Trial; Placebo.

## **Introduction**

The Hamilton Rating Scale for Depression (HAM-D) is the most widely used clinician-rated scale for the assessment of depression severity in patients who were already diagnosed with a depressive disorder [1]. The first version of the HAM-D was originally published by Max Hamilton in 1960 and consisted of 21 items, the HAM-D<sub>21</sub> [2]. The number of citations of this version of the scale exceeds 21.000 on Scopus. Hamilton himself recommended, however, to use only the first 17 items of the HAM-D<sub>21</sub> since the last four symptoms (i.e., diurnal variation, depersonalization/derealization, paranoid and obsessional/compulsive symptoms) were either not considered part of the disease, or they were relatively uncommon, or they were not considered as features related to depression severity [2, 3]. Over the years, a number of modifications of the scale have been proposed and several versions of the HAM-D have been developed and come into use [4, 5].

## **Different versions of the HAM-D**

The main aspects differentiating the various versions of the HAM-D that were developed over the years are the following:

### *Items*

The versions of the HAM-D that are available differ in the number, sequence and wording of items, including the scoring procedure. The items of the most widely used versions of the HAM-D are reported in Table 1. Rosenthal and Klerman [6] proposed the first modified version of the HAM-D, which included an additional item on worthlessness. Paykel et al. [7] introduced substantial modifications with the development of the Clinical Interview for Depression (CID), an expanded version of the HAM-D encompassing 36 items [8, 9]. Kovacs et al. [10] referred to a 24-item version of the HAM-D, the HAM-D<sub>24</sub>, which, in addition to worthlessness, also contained symptoms of helplessness and hopelessness. Gelenberg et al. [11] used an expanded version of the HAM-D, consisting of 27 items, the HAM-D<sub>27</sub>, which covered symptoms of atypical depression. Similarly, Thase et al. [12] introduced another expanded version of the HAM-D to evaluate reverse neurovegetative symptoms such as hypersomnia, increased appetite, and weight gain. Authors also introduced shortened versions of the HAM-D, as pioneered by Bech and his research group [13-15]. They demonstrated the validity of the 6-item version of the HAM-D, the HAM-D<sub>6</sub>, for the assessment of the core (central) symptoms of depression [13-15]. Maier and Philipp [16] reported similar results, since they identified a 6-item version of the HAM-D, which had 5 items in common with the version proposed by Bech et al. [13-15]. In a study from 1993, Gibbons et al. [17] proposed another short version of the HAM-D, including 8 items, which were found to be a unidimensional measure of depression severity.

### *Unstructured and structured versions*

Modalities of administration included: *unstructured* versions with ratings only [2, 3, 11], devoid of interview guides and anchor points; *structured* versions with at least anchor points [18], supplemented by semi-structured [8, 19, 20] or structured interview questions [21-28].

The format of the original version of the HAM-D, the HAM-D<sub>21</sub> [2], was unstructured: Max Hamilton provided only general instructions for rating the individual items of the HAM-D<sub>21</sub>. As he stated [2], the interview process “depends entirely on the skill of the interviewer in eliciting the necessary information”. In other terms, the unstructured versions of the HAM-D, particularly the HAM-D<sub>21</sub> and the HAM-D<sub>17</sub> [2, 3], exclusively rely on expertise and clinical judgment of raters.

The version developed by Bech et al. [18], the HAM-D<sub>23</sub>, explicitly provided item definitions and “operational criteria” (i.e., anchor points) for rating each item of the HAM-D. Anchor points (most of them ranging from 0 to 4) were introduced for detecting the presence and severity of depressive symptoms [18].

In the mid sixties, Eugene S. Paykel started development of the Clinical Interview for Depression [7] to provide a comprehensive assessment of depression. The CID includes item definitions, specific anchor points rated on 7-point scales, and a semi-structured interview, with specified initial questions for each item, which may be modified if circumstances necessitate, and further probing whenever a symptom is present [8, 9]. Morriss et al. [19] published a structured version of the HAM-D with semi-structured interview, which consisted of 17 items and included anchor points for rating the severity and frequency of depressive symptoms. Moreover, Timmerby et al. [20] introduced a structured version of the HAM-D<sub>6</sub>, which included anchor points and interview guides for evaluating the severity of core symptoms of depression.

Further versions of the HAM-D including structured interviews were also developed [21-28] with the purpose of improving inter-rater reliability, as well as individual item reliability. In the appendix of a book on interpersonal psychotherapy, Klerman et al. [29] introduced the first structured version of the HAM-D, the Interview Format for the HAM-D<sub>21</sub>. Miller et al. [21] developed the Modified Hamilton Rating Scale for Depression (MHRSD), a 25-item structured interview guide, which included additional items for the assessment of cognitive and melancholic symptoms of depression and anchor points [30]. Whisman et al. [23] published a 17-item version of the HAM-D with structured interview, the DIS-HRSD, which they integrated with the National Institute of Mental Health (NIMH) Diagnostic Interview Schedule (DIS) [31]. Potts et al. [24] modified the original unstructured version of the HAM-D<sub>17</sub> [3] to be suitable for a structured interview, the SI-HDRS. Janet B.W. Williams [22, 25] published the Structured Interview Guide for the HAM-D<sub>21</sub>, the SIGH-D, which consisted of a set of standard questions and anchor points for rating the frequency and severity of depressive symptoms. Some years later, Williams and colleagues [26] published the Structured Interview Guide for Seasonal Affective Disorders, the SIGH-SAD, a 29-item clinician rated scale, which was specifically developed for the assessment of symptoms of atypical depression (e.g., hyperphagia, hypersomnia). A revised version of the SIGH-D, in which three items were added and anchor points were provided for assessing symptoms of hopelessness, helplessness, and worthlessness, was

developed by Moberg et al. [27]. Williams et al. [28] also introduced the GRID-HAM-D, a structured interview guide, in which the severity and frequency of depressive symptoms are rated separately. The GRID-HAM-D scoring system also included the following modifications: the item content was further clarified and anchor point descriptions were provided with clinical examples at each severity level [28].

## **Reviews on the HAM-D**

A number of reviews were conducted on the HAM-D [1, 20, 32-48]. Hedlund and Vieweg [49] performed the first one in 1979. However, most of these reviews focused on the psychometric properties of the HAM-D [33, 38, 40, 42, 43, 48]. The authors of such reviews did not analyze the different versions of the HAM-D and, most importantly, they did not evaluate the clinimetric properties of these rating scales [38-40, 46, 48]. To the best of our knowledge, only Williams [4] conducted a comprehensive review to evaluate the validity of the different versions of the HAM-D. However, also in this case, the author did not focus on the clinimetric features of these rating scales [4]. Timmerby et al. [20] conducted a systematic review on the clinimetric properties of the HAM-D, in which they compared the HAM-D<sub>6</sub> and the HAM-D<sub>17</sub> without differentiating between unstructured and structured versions. There is, therefore, a need for an updated and comprehensive work, particularly to address the clinimetric properties of the various versions of the HAM-D.

## **From psychometrics to clinimetrics**

It was Alvan R. Feinstein [50-52] who coined the term “clinimetrics” to introduce an innovative approach, which has been defined as the science of clinical measurements [53]. Such a clinically based evaluation method is particularly useful for testing a number of measurement properties (e.g., sensitivity, scalability, clinical validity), which do not find room in the traditional psychometric model [53-55]. Homogeneity of components, as measured by statistical analyses such as Cronbach’s alpha coefficients and factor analyses, has been considered the most important requirement for a psychometric rating scale [53-60]. In the psychometric model, redundant items (i.e. questions which are highly correlated to each other) are needed for ensuring such homogeneity of components [55]. However, the same properties that give a scale a high score for homogeneity may obscure its clinical utility, particularly its ability to detect change [55, 61]. In the clinimetric approach, homogeneity of components is not needed and what matters is the sensitivity of the rating scale, i.e. its ability to discriminate between active therapy and placebo, to differentiate patients from healthy controls, to discriminate between different groups of patients, to differentiate the severity of symptoms (e.g., certain symptoms may be more troublesome or incapacitating than others), to detect clinically relevant changes in drug or psychotherapy trials [53, 55, 62-65]. Such clinimetric properties are particularly important when treatment effects are small and in the evaluation of sub-clinical symptoms [64].

## **Aims**

The present comprehensive review of studies was conducted to describe and evaluate the clinimetric properties of different versions of the HAM-D. The major aim of this review was to identify the versions of the HAM-D that best display the clinimetric properties of reliability, validity, and sensitivity to change in the assessment of depression.

In view of the amount of literature on this topic, the present review cannot be systematic, but will analyze the most representative studies. Since the CID [8] presents substantial differences from the HAM-D and its clinimetric properties have already been analyzed in a previous review study [9], comparisons between the HAM-D scales and the CID will be briefly described in the results section, and then discussed.

## **Methods**

### *Search strategy*

A comprehensive search of the literature was conducted on the following databases: MEDLINE, Scopus, Web of Science, and PubMed. Each database was searched from inception to July 23, 2019. A manual search of the literature was also performed and reference lists of the retrieved articles were examined for further studies not yet identified. The following search terms were used: “Hamilton Rating Scale for Depression”, “HRSD”, “Hamilton Depression Rating Scale”, “HAM-D”. They were combined using the Boolean operator “OR”.

### *Eligibility criteria*

We selected and analyzed only those studies, which focused on the clinimetric properties of the HAM-D. To be included in the review, studies had to meet the following inclusion criteria:

1. English-language article published in a peer-reviewed journal.
2. The study was published as a full-text article.
3. The article was an original study (e.g., research article, meta-analysis).
4. The study evaluated the clinimetric properties of the HAM-D or used a clinimetric approach to analyze the clinical utility of this rating scale.

### *Study selection procedure*

The first two authors (D.C. and C.P.) independently performed the search, screened titles and abstracts, selected studies, evaluated the full-text of articles appearing potentially relevant, and extracted data from studies meeting the eligibility criteria. In case of disagreement, a consensus was reached through discussion with the last author (J.G.).



## Results

The initial search yielded a total of 35,473 articles, but only those studies, which best displayed the clinimetric properties of the HAM-D, were included in the review. Accordingly, the inter-rater and test-retest reliability, the discriminant validity, the sensitivity to change, the scalability, and the concurrent validity of the various versions of the HAM-D will be examined in detail.

### Inter-rater reliability

Trajković et al. [46] conducted a meta-analytic study to examine the inter-rater reliability of the HAM-D. They found a pooled mean ICC of 0.92, indicating an excellent level of inter-rater reliability. Unfortunately, they did not discriminate between the different versions of the HAM-D [46].

#### *Unstructured versions*

Max Hamilton [2] examined the inter-rater reliability of his unstructured version of the HAM-D<sub>17</sub>. He found that the correlation between two raters who independently evaluated 10 depressed patients was 0.84 [2]. He used skilled interviewers who were experienced in the use of the HAM-D<sub>17</sub> [2]. In a study performed in 1975, Bech et al. [13] adopted the same procedure: two experienced psychiatrists independently scored the unstructured HAM-D<sub>17</sub>. The inter-rater reliability was found to be excellent with a Spearman's correlation coefficient of 0.94 [13]. O'Hara and Rehm [30] reported similar results. They demonstrated that the skill level or expertise of the rater positively affected the inter-rater reliability of the HAM-D<sub>17</sub> [30]. They indeed showed that the agreement was high (intra-class correlation coefficient – ICC of 0.91) for expert raters but low for novice raters (ICC = 0.76) [30]. Many subsequent studies confirmed this trend: the inter-rater reliability of the unstructured versions of the HAM-D, particularly that of the HAM-D<sub>21</sub> [66] and of the HAM-D<sub>17</sub>, was largely influenced by the clinical experience of raters: the greater the expertise of raters, the higher the inter-rater reliability [67-81].

#### *Unstructured versus structured versions*

Compared to the unstructured HAM-D<sub>17</sub>, the inter-rater reliability of those versions including anchor points and interview guides was significantly higher independently if interviewers were experienced clinicians or novice raters [19, 21, 22, 27, 82]. Potts et al. [24] evaluated the inter-rater reliability of their structured version of the HAM-D<sub>17</sub>, the SI-HDRS, and found that the level of agreement between inexperienced raters was excellent: ICC of 0.92. Moberg et al. [27] compared the unstructured HAM-D<sub>24</sub> with their structured version of this rating scale. They found that the structured HAM-D<sub>24</sub> produced significantly higher levels of inter-rater reliability than the unstructured one [27]. Williams et al. [28] conducted a similar study. They compared the inter-rater reliability of the GRID-HAM-D with that of the unstructured version of the HAM-D<sub>17</sub> and found that the ICC (0.78) for the unstructured HAM-D<sub>17</sub> was significantly lower ( $p = 0.001$ ) than the ICC (0.95) for the structured GRID-HAM-D [28]. Tabuse et al. [82]

provided further support to the inter-rater reliability of the GRID-HAM-D. They showed that this structured version of the HAM-D displayed excellent inter-rater reliability for both inexperienced and experienced raters before and after training [82].

#### *Structured versions with anchor points only*

The majority of studies examined the inter-rater reliability of the structured version of the HAM-D, which used the anchor points introduced by Bech et al. [18]. Kørner et al. [83] conducted one of the first studies in this regard. They found an ICC of 0.83, demonstrating a high level of agreement between raters [83]. Koenig et al. [84] and Fuglum et al. [85] reported similar results. The former authors [84] found inter-rater correlations between 0.93 and 0.98, while the latter researchers [85] showed an ICC of 0.81, in both cases indicating a high level of inter-rater reliability. Bent-Hansen and Bech [86] performed a similar study and found an ICC of 0.95, indicating a high level of inter-rater reliability.

#### *Structured versions with semi-structured interview*

Paykel et al. [7] examined the inter-rater reliability of the CID and found a Pearson's correlation coefficient of 0.81, indicating a satisfactory level of inter-rater agreement. The inter-rater reliability of the CID was evaluated in other studies, yielding similar results: the agreement between raters was high, with mean correlation coefficients ranging from 0.81 [87] to 0.82 [88]. Morriss et al. [19] evaluated the inter-rater reliability of a structured version of the HAM-D<sub>17</sub> and found ICCs ranging from 0.89 to 0.96, indicating excellent inter-rater reliability.

#### *Structured versions with structured interview*

Many studies examined the inter-rater reliability of the different structured versions of the HAM-D. Whisman et al. [23] showed that the inter-rater reliability of their structured version of the HAM-D, the DIS-HRSD, was satisfactory. Specifically, they found an ICC of 0.84, indicating a high level of agreement between raters [23]. Akdemir et al. [89] examined the inter-rater reliability of the SIGH-D, the structured interview guide published by Williams [22]. They found Pearson correlation coefficients ranging from 0.87 to 0.98, indicating a high level of agreement between raters [89]. Rohan et al. [90] evaluated the inter-rater reliability of the SIGH-SAD, another structured interview developed by Williams and her research group [26]. They found ICC coefficients (ICCs) ranging from 0.92 to 0.96, indicating an excellent level of inter-rater reliability [90]. As to the inter-rater reliabilities of individual items of the HAM-D structured versions, authors reported similar results [21, 22, 25, 27]. The inter-rater reliabilities of the individual items included in the structured versions of the HAM-D were significantly higher than those obtained in studies, in which anchor points and interview questions were not used [21, 22, 25, 27].

## **Test-retest reliability**

Test-retest reliability refers to the ability of an assessment instrument to produce the same results over time, while it is assumed that the clinical dimension under examination has remained unchanged [91, 92] In clinimetrics, such a measurement property is not considered to be as important as other clinimetric properties since the rating scale is primarily intended to be used for detecting treatment changes [9, 55].

### *Unstructured versions*

Cicchetti and Prusoff [71] were among the first authors to examine the test-retest reliability of an unstructured 22-item version of the HAM-D. They found poor to fair levels of test-retest reliability for most of the HAM-D items [71]. Craig et al. [72] evaluated the test-retest reliability of the unstructured version of the HAM-D<sub>17</sub> in a small sample of inpatients with schizophrenia. Using raters who had previously received training in the use of the HAM-D<sub>17</sub>, they showed high test-retest reliability for the HAM-D<sub>17</sub> total score, with a correlation coefficient of 0.65 [72]. However, analyzing the individual items of the HAM-D<sub>17</sub>, poor test-retest reliability was found for items on guilt feelings, somatic symptoms (gastrointestinal), genital symptoms, hypochondriasis, and weight loss [72].

### *Structured versions*

Williams [22] demonstrated that the use of her structured version of the HAM-D, the SIGH-D, significantly improved the test-retest reliability for most of the SIGH-D items. More specifically, compared to the study by Cicchetti and Prusoff [71], Williams [22] showed that all but three (late insomnia, psychomotor retardation, and agitation) of the 21 SIGH-D items had better test-retest reliability. Using their structured version of the HAM-D<sub>17</sub>, the SI-HDRS, Potts et al. [24] reported similar results: all but four items (i.e., loss of weight, insight, psychomotor agitation, and psychomotor retardation) of the 17 SI-HDRS items had satisfactory test-retest reliability. Other studies demonstrated that the different structured versions of the HAM-D produced uniformly higher test-retest reliabilities than the unstructured ones. Akdemir et al. [89] investigated the test-retest reliability of the SIGH-D and found a correlation coefficient of 0.85, indicating high test-retest reliability for the SIGH-D total scores. They also showed that the use of this structured version improved the test-retest reliability for all but one (i.e., loss of weight) of the 17 HAM-D items [89]. Shankman and Klein [93] examined the test-retest reliability of the MHRSD, the structured interview of the HAM-D developed by Miller et al. [21]. They found that the MHRSD had excellent test-retest reliability, with an ICC of 0.96 [93]. Williams et al. [28] tested the test-retest reliability of their structured GRID-HAM-D. They showed satisfactory level of test-retest reliability, with an ICC of 0.81 [28].

## **Discriminant validity**

According to the clinimetric approach, the validity of a rating scale is established using the global assessment of the experienced clinician as the main index of validity [55, 91, 94]. The aim is to determine whether the items included in the rating scale reflect the clinician's judgment of severity of the clinical condition under assessment [91, 95].

Chipman and Paykel [87] found that specific individual items of the CID correlated with the clinician's global assessment of depression severity. They indeed reported that patients rated as more severely depressed by clinicians were those reporting higher scores on the following items of the CID: psychomotor retardation, depressive delusions, agitation, guilt, initial insomnia, hopelessness, suicidal tendencies, verbal complaint of depressed feelings, observed appearance of depression, and less short-term reactivity of mood [87]. Bech et al. [13] conducted a similar study for evaluating the validity of the unstructured version of the HAM-D<sub>17</sub>. They demonstrated that only 6 of the 17 items of the HAM-D (those included in the HAM-D<sub>6</sub>) reflected the clinician's evaluation of depression severity [13]. More specifically, only the items of being depressed, having guilt feelings, experiencing lack of interest and fatigue, displaying psychomotor retardation, and suffering from psychic anxiety corresponded to symptoms used by experienced clinicians to formulate a global evaluation of depression severity [13]. Therefore, the total score of the HAM-D<sub>6</sub> was found to be strongly associated with the clinician's global impression of depression severity [13]. Further studies showed that the HAM-D<sub>6</sub> sensitively captured core symptoms of depression better than the HAM-D<sub>17</sub>, that actually covers a mixture of anxiety and depressive symptoms, including side effects of pharmacological treatments such as nausea, weight gain, and sexual dysfunction [13-15, 20, 67, 96-100].

### *Depression severity*

As to studies evaluating the ability to discriminate between different groups of patients suffering from the same illness, Carroll et al. [101] were among the first authors to demonstrate that the total score of the unstructured version of the HAM-D<sub>17</sub> sensitively differentiated severely depressed inpatients from moderately and mildly depressed outpatients. In other words, they showed that inpatients with severe depression scored significantly higher on the HAM-D<sub>17</sub> than the other two groups of depressed patients [101]. Knesevich et al. [67] reported similar findings. Using the global judgment of experienced clinicians, they allocated a small sample of 26 depressed patients into four severity groups: none, mild, moderate, and severe [67]. Then, they showed that the total score of the unstructured version of the HAM-D<sub>17</sub> sensitively differentiated between the four different levels of depression severity [67]. Thase et al. [102] showed that the unstructured version of the HAM-D<sub>17</sub> sensitively differentiated patients with endogenous depression from those with nonendogenous depression. They found that patients with endogenous depression scored significantly higher on the HAM-D<sub>17</sub> than patients with nonendogenous depression [102]. Zheng et al. [77] used the Global Assessment Scale (GAS) developed by Endicott et al. [103] to evaluate depression severity

and then they tested the discriminant validity of the unstructured version of the HAM-D<sub>17</sub>. They showed that the HAM-D<sub>17</sub> total score sensitively discriminated between different levels of depression severity [77]. They indeed demonstrated that patients with higher HAM-D<sub>17</sub> scores were likely to be more severely disabled according to the GAS [77].

#### *Depressed patients versus healthy controls*

As to studies examining the ability of the HAM-D to differentiate patients from healthy subjects, Ganchrow et al. [104] showed that the unstructured HAM-D<sub>17</sub> sensitively discriminated patients with depression from controls. Fava et al. [105] conducted a similar study, in which they demonstrated that only 17 items of the unstructured version of the HAM-D<sub>21</sub> (the first 16 questions and the item on diurnal variation of mood) sensitively discriminated depressed patients from healthy controls. Rehm and O'Hara [73] reported similar results. The total score of the unstructured version of the HAM-D<sub>17</sub> sensitively differentiated depressed patients from healthy controls [73]. However, they showed that four items (i.e., agitation, gastrointestinal symptoms, loss of insight, and weight loss) failed to discriminate between depressed patients and controls [73].

#### *Different groups of patients*

As to studies analyzing the ability of the various versions of the HAM-D to discriminate between different groups of patients, Rush et al. [106] found that the total score of the unstructured version of the HAM-D<sub>17</sub> sensitively differentiated patients with major depression from those with other psychiatric diagnoses (e.g., bipolar disorder, schizophrenia, generalized anxiety disorder, panic disorder).

Carneiro et al. [107] showed that only four items of the unstructured HAM-D<sub>17</sub> (i.e., insomnia late, general somatic symptoms, hypochondriasis, and insight) sensitively discriminated depressed patients from bipolar I patients.

#### *Cut-off scores*

In the clinimetric approach, to be considered clinically meaningful, cut-off scores should be tested using the global judgment of the experienced clinician as the gold standard [91, 108].

Using the judgment of the experienced clinician as the main index of validity, Zimmerman et al. [109] established score ranges for the HAM-D<sub>17</sub> reflecting different levels of depression severity. They found that a score ranging from 8 to 16 corresponded to mild depression, while a score range from 17 to 23 reflected moderate depression [109]. They also showed that a score  $\geq 24$  on the HAM-D<sub>17</sub> was indicative of severe depression [109]. Using the same approach (i.e., global impression of the experienced clinician as the main index of validity), Kyle et al. [100] established cut-off scores for remission in the different versions of the HAM-D. Cut-off scores, which were found to be clinically valid indicators of remission, were the following: a score of  $< 5$  for the HAM-D<sub>6</sub>, a cut-off score of less than 8 for the HAM-D<sub>17</sub>, a score of  $< 9$  for

the HAM-D<sub>21</sub>, and a cut-off score of less than 10 for the HAM-D<sub>24</sub> [100]. Bobo et al. [99] applied the same clinimetric approach but, in this case, they focused on the clinician's global impression of improvement to identify cut-off scores indicative of a clinically significant level of change in the unstructured HAM-D<sub>17</sub> and HAM-D<sub>6</sub>. They found that a clinician's global evaluation of improvement was associated to a reduction of 11 points in the HAM-D<sub>17</sub> scores, corresponding to a percent reduction of 50 to 57% in the HAM-D<sub>17</sub> scores [99]. As to the HAM-D<sub>6</sub>, the clinician's global impression of improvement was found to be associated to an absolute reduction of 7 points in the HAM-D<sub>6</sub>, corresponding to a percent reduction of 57% to 63% [99].

#### *Discriminant validity of the HAM-D<sub>17</sub> compared to the CID*

A number of studies compared the discriminant validity of the HAM-D<sub>17</sub> with that of the CID in general practice. In a study by Freeling et al. [110, 111], patients whose major depression had gone unrecognized by their physicians appeared to be significantly less severely depressed on the CID, but not on the HAM-D<sub>17</sub>, than those whose depression had been recognized. On both the HAM-D<sub>17</sub> and the CID patients with unrecognized depression showed less evidence of overt depressed mood. On the HAM-D<sub>17</sub> they showed greater lack of insight, whereas on the CID they were less obviously depressed at interview based on their appearance. Patients with unrecognized depression had also higher scores on the CID reactivity to social environment and distinct quality of mood. When depressed patients receiving a new prescription of an antidepressant in general practice (GP) were compared with those given other treatments, and with antidepressant-treated psychiatric outpatients [112], the mean HAM-D<sub>17</sub> and CID depression scores were considerably higher in the outpatients than in the two GP samples. Significant differences were found also between the two GP samples on both scales, with higher scores for antidepressant-treated GP patients compared to those receiving other treatments. The CID provided a detailed description of specific symptom patterns for each subgroup. Differences between GP female patients with recognized and unrecognized depression in their symptom ratings were found for two individual items of the CID (i.e., tiredness and distinct quality of depressed mood), but not on the HAM-D<sub>17</sub> [113].

Further, significantly slower improvement 2 weeks after admission was detected with the CID in depressed inpatients with comorbid personality disorders (PD) compared to those without, even though differences did not reach significance on the HAM-D<sub>17</sub> [114]. In another study [115], both the HAM-D<sub>17</sub> and the CID sensitively discriminated between acutely and remitted depressed patients.

#### **Sensitivity to change**

The differentiation between an active drug and placebo [54, 63, 64, 116, 117] or between a specific psychotherapeutic treatment and attention placebo or clinical management [65] is particularly important when treatment effects are small and in the setting of subclinical symptoms. Kellner and Sheffield [62] used the term "sensitivity" to describe this clinimetric property.

### *Comparison with observer-rated scales*

Montgomery and Åsberg [118] were among the first authors who criticized the unstructured version of the HAM-D<sub>17</sub> [2] for being poorly sensitive in detecting change during treatment. They developed a new rating scale, the Montgomery-Åsberg Depression Rating Scale (MADRS) [118], which was specifically designed to be more sensitive than the HAM-D<sub>17</sub>. Actually, when Khan et al. [37] analyzed records of 208 depressed patients, who participated in eight randomized, placebo-controlled, double-blind clinical trials between 1996 and 2000, they found that the MADRS was just as sensitive as the HAM-D<sub>17</sub> in differentiating between antidepressants and placebo. Studies also showed that the HAM-D<sub>6</sub>, the short version of the HAM-D introduced by Bech et al. [13-15], was more sensitive than the MADRS [119-125]. The HAM-D<sub>6</sub>, but not the MADRS, was sensitive to the superior antidepressant efficacy of mirtazapine over trazodone [119]. In another study testing the antidepressive effects of hypericum, Lecrubier et al. [122] showed that the HAM-D<sub>6</sub>, but not the MADRS, sensitively discriminated between active drug and placebo. Similarly, Liebowitz et al. [124] found that the HAM-D<sub>6</sub>, but not the MADRS, detected the antidepressant superiority of desvenlafaxine over placebo. The sensitivity to change of the HAM-D<sub>6</sub> has been tested in a number of other clinical trials [121, 123, 126-132].

Helmreich et al. [133] compared the sensitivity of the unstructured HAM-D<sub>17</sub> with that of the 28-item clinician version of the Inventory of Depressive Symptomatology, the IDS-C<sub>28</sub>, a relatively new rating scale, which was developed by Rush and his research group [106, 134] for evaluating signs and symptoms of depression. Using data from 340 patients in a 10-week randomized, placebo-controlled trial comparing the effectiveness of sertraline and cognitive-behavioral therapy, they found that the IDS-C<sub>28</sub> was more sensitive than the HAM-D<sub>17</sub> in detecting small changes in depression symptomatology over the treatment course [133]. Liu et al. [135] compared the sensitivity of the unstructured HAM-D<sub>17</sub> with that of the 16-item clinician version of the Quick Inventory of Depressive Symptomatology, the QIDS, another rating scale, which was developed by Rush and his research group [136, 137]. Evaluating depression at baseline and 6 weeks later, they found that the QIDS and the HAM-D<sub>17</sub> were equally sensitive to change of depressive symptoms [135].

### *Sensitivity of HAM-D<sub>6</sub> compared to that of HAM-D<sub>17</sub>*

Evaluating the antidepressant effects of citalopram, Østergaard et al. [138] found that the HAM-D<sub>6</sub> was more sensitive than the HAM-D<sub>17</sub> in capturing the antidepressant effects of citalopram. In another study, Bech et al. [139] found that the HAM-D<sub>6</sub>, but not the HAM-D<sub>17</sub>, was sensitive to the superior antidepressant efficacy of bupropion over buspirone. Many other studies reported similar results: the HAM-D<sub>6</sub>, but not the HAM-D<sub>17</sub>, was found to sensitively differentiate between different antidepressant effects [120-122, 124-126, 140-146].

### *Differentiating active treatment from placebo*

Studies showed that the HAM-D<sub>6</sub>, but not the HAM-D<sub>17</sub>, sensitively discriminated between active drug and placebo [139, 147-149]. In Chouinard et al. [147], brofaromine was statistically ( $p < 0.050$ ) superior to placebo on the HAM-D<sub>6</sub>, but did not significantly differ from placebo on the HAM-D<sub>17</sub>. Fabre et al. [126] obtained similar results. Using the HAM-D<sub>6</sub>, they showed that sertraline was significantly superior to placebo at all three doses (i.e., 50, 100, and 200 mg daily) [126]. This finding was not replicated when they used the HAM-D<sub>17</sub> [126]. In a subsequent study [148] the HAM-D<sub>6</sub>, but not the HAM-D<sub>17</sub>, sensitively discriminated active treatment from placebo. More specifically, using the HAM-D<sub>6</sub>, Feiger et al. [148] found that selegiline was statistically ( $p < 0.01$ ) superior to placebo in decreasing symptoms of major depression.

### *Comparison with self-rating scales*

Studies compared the sensitivity of the various versions of the HAM-D with that of self-reported questionnaires. Carroll et al. [101] were among the first authors who compared the sensitivity of the unstructured version of the HAM-D<sub>17</sub> with that of the Zung Self-Rating Depression Scale. They showed that the HAM-D<sub>17</sub>, but not the Zung Self-Rating Depression Scale, sensitively discriminated severely depressed inpatients from outpatients with moderate or mild depression [101]. Edwards et al. [150] conducted a meta-analysis, in which they compared the sensitivity of the unstructured version of the HAM-D<sub>17</sub> with that of the Beck Depression Inventory (BDI), one of the most widely used questionnaires for the assessment of self-reported symptoms of depression [151]. They found that the HAM-D<sub>17</sub> was more sensitive to change than the BDI [150]. Other studies reported similar results: the HAM-D<sub>17</sub>, but not the BDI and the Zung Self-Rating Depression Scale, was highly sensitive to change [32, 152].

### *Comparison with the CID*

The HAM-D<sub>17</sub> and the CID were used in a placebo-controlled trial of amitriptyline among depressed patients in GP [153, 154]. Several individual items of the CID showed the superiority of amitriptyline over placebo over 6 weeks of treatment, whereas only four items of the HAM-D<sub>17</sub> (i.e., depressed mood, guilt, early and late insomnia) displayed significant drug-placebo differences [155]. In a study on relapse prevention with cognitive therapy (CT) in residual depression [156, 157], similar non-significant lower HAM-D<sub>17</sub> and CID scores were found at 1-year follow-up in the CT + clinical management (CM) group compared to the CM only group. Differences between groups were most marked at the end of treatment and the next 6 months, and were not fully lost until 3 ½ years after the end of CT [158]. Significant time by group interactions were found at 1-year follow-up for the CID depression score and two individual items (i.e., guilt and self-esteem, and hopelessness and pessimism), but not for the HAM-D<sub>17</sub> [157].



## Scalability

In the clinimetric approach, it is particularly important to test whether the items of the rating scale reflect one single dimension of severity [20, 55, 91, 159]. More specifically, item response theory (IRT) models (i.e., Rasch and Mokken analyses) are required to determine the level of scalability of the rating scale [20, 55, 91, 159]. The Rasch analysis is the parametric version of IRT models [160], while the Mokken analysis is the corresponding non-parametric version [91, 161]. In the Rasch model, the scalability is evaluated by using a number of fit indices such as the differential item functioning (the difficulty level of each item), the invariant item ordering (the expected score for an “easy” item is always higher than the expected score for a “hard” item) and the local independence of items (the probability of a positive score on an item should not depend from a positive score on any other item) [91, 160, 162]. When the data meet these criteria, the Rasch model assumes that both the respondent’s ability (e.g., his/her level of depression) and the degree of clinical information (e.g., the severity of depression) measured by the item are evaluated on the same scale [91, 160, 162]. Clinically, it means demonstrating that all items of the rating scale are multidimensional, i.e. they measure different symptoms but of the same clinical dimension [55, 91, 163]. In the Mokken analysis, the scalability is evaluated by using the Loevinger’s coefficient [164]. Such a clinimetric coefficient is an expression of the extent to which each item of the rating scale covers a specific (i.e., unique) level of symptom severity of an underlying clinical dimension [55, 91]. According to Mokken [161], a Loevinger’s coefficient  $\geq 0.30$  indicates not only that items are not redundant but also that the total score of the rating scale is a statistically sufficient and clinically valid measure of the severity of the clinical condition under assessment [91]. The clinimetric property of scalability is thus important to differentiate the various versions of the HAM-D.

### *Scalability of the different versions of the HAM-D*

Bech and his research group conducted a number of studies in which they analyzed the scalability of the various versions of the HAM-D [14, 20, 96, 98, 100, 165-172]. In these studies, including a recent paper by da Silva et al. [173], the HAM-D<sub>6</sub> was found to have an excellent level of scalability with Mokken coefficients ranging from 0.42 [172] to 0.65 [169]. Studies using the Rasch analysis further confirmed the scalability of the unstructured version of the HAM-D<sub>6</sub> [96, 165, 170]. The other HAM-D versions, which were found to have an acceptable level of scalability, were the unstructured HAM-D<sub>24</sub> [100] and the CID [174]. Using the Rasch analysis, Bech, Paykel and colleagues [174] demonstrated that the CID contains valid subscales for the assessment of affective disorders such as depression, anxiety, and apathy.

Conflicting results were obtained for the scalability of the unstructured version of the HAM-D<sub>17</sub>. Mokken coefficients ranging from 0.24 [172] to 0.35 [169] were found, indicating that the HAM-D<sub>17</sub> is a multidimensional rating scale. Similarly, Bobes et al. [98] and Kyle et al. [100] questioned the scalability of the unstructured version of the HAM-D<sub>21</sub>. They found Mokken coefficients ranging from 0.29 [100] to 0.30 [98], indicating that the original version of the HAM-D, the HAM-D<sub>21</sub>, is a multidimensional rating scale.

## Concurrent validity

A high correlation is often regarded as evidence that two rating scales measure the same clinical factor [53]. However, a high correlation does not indicate similar clinical validity [53, 91, 108]. Two rating scales may have a common content, which insures a high positive correlation, but they may display differential sensitivity [53, 64]. The concurrent validity of the different versions of the HAM-D has been widely examined, especially by using self-reported questionnaires.

### *Different versions of the HAM-D compared to self-rating scales*

Prusoff et al. [175] were among the first authors to test the concurrent validity of the unstructured version of the HAM-D<sub>17</sub>. Using a sample of 200 depressed patients, they found that the HAM-D<sub>17</sub> had low correlations with self-rating scales at baseline (i.e., during the acute episode of illness), while the correlations were significantly higher at follow-up, when patients improved with treatment [175]. The authors concluded that, compared to self-rating scales, the HAM-D<sub>17</sub> was a better measure of depression severity [175]. Focusing on a sample of 40 depressed outpatients and 40 healthy controls, Fava et al. [105] examined the concurrent validity of the unstructured version of the HAM-D<sub>21</sub> using two self-rating scales, the Symptom Rating Test (SRT) [62] and the Symptom Questionnaire (SQ) [176, 177]. They found that there were similar product-moment correlations (ranging from 0.65 to 0.72) between the HAM-D<sub>21</sub> and SRT and the SQ [105]. They concluded that low correlations, such as those found by Prusoff et al. [175], might be due to the specific self-rating scales that were administered rather than to substantial differences between clinician-rated and patient-reported scales [105]. Gottlieb et al. [75] compared the HAM-D<sub>21</sub> with the Zung Self-Rating Depression Scale [178] and showed that in the sub-group of patients with mild Alzheimer's disease (AD) there was a statistically significant correlation between the two rating scales ( $r = 0.49$ ). However, in the sub-group of patients with severe AD, there was no correlation between the HAM-D<sub>21</sub> and the Zung Scale [75]. The authors concluded that self-reported questionnaires are of questionable clinical utility, particularly in patients with advanced AD [75]. Studies evaluated the concurrent validity of the different versions of the HAM-D using also the BDI [23, 73, 89, 179-183]. Rehm and O'Hara [73] found that there was a statistically significant correlation between the total score of the unstructured version of the HAM-D<sub>17</sub> and the BDI ( $r = 0.73$ ). Whisman et al. [23] found that the DIS-HRSD and the unstructured version of the HAM-D<sub>17</sub> exhibited similar correlations to the BDI. Akdemir et al. [89] tested the concurrent validity of the SIGH-D, the structured interview version of the HAM-D, which was published by Williams [22]. Using the total sample of 94 depressed patients, they showed that there was a significant but moderate ( $r = 0.48$ ) correlation between the SIGH-D and the BDI [89]. However, when they analyzed only the sub-group of patients with severe depression, they found that this correlation was no longer statistically significant [89]. The authors concluded that the SIGH-D was clinically superior to the BDI, particularly in the assessment of depression severity [89].

### *Concurrent validity of the HAM-D<sub>17</sub> compared to the CID*

Authors compared the concurrent validity of the HAM-D<sub>17</sub> with that of the CID [184-186], and moderately high correlations were found for the CID depression score, whereas the CID anxiety score correlated to some extent with the HAM-D<sub>17</sub>, reflecting the inclusion of several anxiety items in the latter. Correlations of individual items of the HAM-D<sub>17</sub> and the CID were also examined [187], and were high, except for depressed feelings, which in the CID are rated on intensity on questioning whereas in the HAM-D<sub>17</sub> they are rated for the degree to which they dominate the verbal and nonverbal content of the interview, including observed appearance. Highly significant correlations between the HAM-D<sub>17</sub> and the CID were found also in another study of patient with major depressive disorder [115].

### **Discussion**

Bagby et al. [38] and Zimmerman et al. [40] concluded their reviews that it was time to discard the HAM-D and to embrace a new gold standard for the assessment of depression. Considering the HAM-D a flawed measure, they [38, 40] proposed to adopt psychometrically superior rating scales such as the MADRS [118] and the clinician version of the IDS [106, 134]. The MADRS was specifically developed to be more sensitive to change than the HAM-D [118]. However, the available literature [37] runs counter this assumption, particularly when the HAM-D<sub>6</sub> is used [119, 122, 124]. In fact, it has been shown that the HAM-D<sub>6</sub>, but not the MADRS, sensitively discriminated between active treatment and placebo [122, 124]. Similar considerations apply also to the clinician version of the IDS [168]. In a clinimetric reanalysis of the STAR\*D study, Bech et al. [168] found that the IDS was a poorly sensitive and multidimensional measure of depression severity.

In their reviews, Bagby et al. [38] and Zimmerman et al. [40] criticized other aspects of the HAM-D. They considered the differential item weight as one of the most important limitations of the HAM-D [38, 40]. The evidence that certain HAM-D items contributed more to the total score than others clashed with the psychometric assumption of homogeneity of items [38, 40]. In the psychometric model, to be included in a rating scale, all items have to display the same clinical weight [54, 57, 59, 188]. In the clinimetric approach, however, the differential item weighting is not a disadvantage, but a basic requirement for rating scales [55, 91]. Accordingly, not all items carry the same clinical weight, and major and minor symptoms can be differentiated [54, 55, 59, 91]. Using the HAM-D items, Bech et al. [14] found that symptoms such as depressed mood, loss of interest, and tiredness occurred in both the mildly and the more severely depressed patients, while symptoms such as guilt feelings and psychomotor retardation were only present in the more severely depressed patients. In other words, they demonstrated the clinical utility of differential weighting of HAM-D items, which can be used to sensitively distinguish major depression from moderate to mild depression [14]. Using the HAM-D, Bech et al. [14] also showed that symptoms of depressed mood, loss of

interest, and tiredness were more prevalent and occurred before the onset of more severe symptoms of depression.

Another criticism raised by Bagby et al. [38] and Zimmerman et al. [40] was that the HAM-D items did not cover the DSM diagnostic criteria for major depression. As noted by Zimmerman et al. [40], such an incomplete coverage of the psychiatric classification criteria for depression significantly limited the utility of the HAM-D as a diagnostic measure.

The introduction of diagnostic criteria for the identification of psychiatric syndromes has considerably decreased the variance due to different assessors and the use of inferential criteria rather than direct observation [189]. The diagnostic criteria are particularly helpful in setting a threshold for conditions worthy of clinical attention. Accordingly, the diagnostic criteria for a major depressive disorder identify a syndrome, which may be responsive to antidepressant drugs. At least 5 of a set of 9 symptoms should be present (and 1 should be either depressed mood or loss of interest). However, according to the psychometric model, all items are weighed the same, unlike in clinical medicine, where major and minor symptoms are often differentiated (e.g., Jones criteria for rheumatic fever). As a result, a patient with severe and pervasive anhedonia, incapacitating fatigue and difficulties concentrating which make him unable to work would not be diagnosed as suffering from a major depressive disorder, despite the clinical intuition of potential benefit from pharmacotherapy. This diagnosis could be performed in a patient who barely meets the criteria for 5 symptoms. The hidden conceptual model is psychometric: severity is determined by the number of symptoms, not by their intensity or quality, to the same extent that a score in a depression self-rating scale depends on the number of symptoms that are scored as positive [54, 59].

Bech [91, 190, 191] also showed that rating scales only covering DSM and ICD diagnostic criteria were found to be poorly sensitive to sub-threshold symptoms of depression. The Hamilton scales, particularly the HAM-D<sub>6</sub> and the CID subscales are not diagnostic instruments but rating scales measuring depression severity [1-3]. As has been stated by Corruble and Hardy [192] in this regard: "...the scale should not be compared to DSM-IV criteria because the two measures have different objectives; i.e., the Hamilton depression scale assesses depression severity in depressed patients, and the DSM-IV defines a diagnosis of major depression". Nevertheless, such rating scales can be used in combination with current psychiatric classification systems (e.g., DSM-5). Reaching a concordance between diagnostic criteria and the dimensional evaluation of rating scales is the ultimate goal to achieve for providing a comprehensive assessment of the patient's clinical condition [55, 190, 191]. Therefore, clinician-reported scales, such as the HAM-D<sub>6</sub> and the CID, and DSM or ICD diagnostic criteria are actually not conflicting but convergent evaluation methods [191], which should be used jointly to perform adequate clinical assessment of depression.

Bagby et al. [38] also criticized the reliability of the HAM-D. Particularly, they concluded that the inter-rater and test-retest reliability coefficients were weak for most of the HAM-D items. Actually, reliability was found to be excellent when the structured versions of the HAM-D were used [7, 22, 84-86,

89]. These findings are in line with a recent meta-analysis [46] demonstrating excellent inter-rater reliability for most of the HAM-D individual items.

As noted by Carroll [193], the endurance of the various versions of the HAM-D is remarkable and their utility in the clinical process of assessment of depression severity has been clearly demonstrated.

## Conclusions

The findings of this review indicate that the HAM-D is a valid and sensitive clinimetric index, which should not be discarded in view of obsolete and not clinically relevant psychometric criteria. It is, however, important to note that the various versions of the HAM-D, including the CID, entail different clinimetric properties. Simple reference to the unstructured versions of the HAM-D [2, 3] is no longer acceptable. The choice of the most adequate version depends on a number of clinical factors: raters' clinical experience and their level of training and expertise in the use of the HAM-D, study aims and design, clinical characteristics of the population under examination. In other terms, there is not a unique version of the HAM-D, which best applies to all clinical situations, but investigators are asked to select the version that best fits with the aims of the study and to report appropriate reference details in the registration and publication of the study protocol. Further, in performing meta-analyses, assembling data that derive from different versions of the HAM-D may add to other variables to yield substantial heterogeneity, a major drawback of the meta-analytic method [194, 195]. Similar considerations may apply to the use of the HAM-D in network analysis [196, 197].

A few indications emerge from the literature. The total score of the HAM-D<sub>17</sub>, and not that of the HAM-D<sub>21</sub>, should be used for sensitively discriminating between different levels of depression severity. If the aim of the investigation is to discriminate between active treatment and placebo, the HAM-D<sub>6</sub> and the CID should be considered. The CID, in particular, appears to be indicated when differences are expected to be small and/or when dealing with mild or subclinical symptoms [9, 198]. In fact, the use of 7-point Likert scales, as reported by Bech and colleagues [174], makes the CID particularly suitable to sensitively capture the relatively milder or subsyndromal symptoms of depression. By contrast, there appears to be no valid justification for using the MADRS, which was found to be less sensitive than the HAM-D<sub>6</sub> or as sensitive as the HAM-D<sub>17</sub> to changes with treatment [20, 37, 119, 122, 124].

Regardless of the number of items, unstructured versions of the HAM-D [2, 3] should be avoided, while the use of structured versions in clinical trials appears to be mandatory to ensure inter-rater reliability. In view of its clinimetrics properties [9, 174], the CID should be considered as the best structured version of the HAM-D and it can be regarded as the gold standard for the assessment of depression severity. The use of the structured versions of the HAM-D can be supplemented by other indices based on the clinimetric principle of incremental validity [163, 189]. Accordingly, each distinct aspect of psychological measurement should deliver a unique increase in information in order to qualify for inclusion and this should be applied to the selection of instruments in a clinical trial. Several highly redundant scales are often used under the

misguided assumption that nothing will be missed. On the contrary, violation of the concept of incremental validity only leads to conflicting results. Structured versions of the HAM-D, in particular the CID, may thus be used in conjunction with self-rating scales for depression and anxiety that are likely to improve incremental validity. Kellner's Symptom Questionnaire, in view of the data that are available and its high sensitivity to change [176, 177], appears to be very suitable for this purpose. Transdiagnostic clinimetric indices that may assess other dimensions, such as psychological well-being and euthymia [163, 199-201] or mental pain [202, 203], may also add valuable information.

### **Disclosure Statement**

All authors have no conflicts of interest to declare.

### **Funding sources**

None.

### **Author Contributions**

All Authors conceived the project. D.C and C.P. performed the searches and collected the data. All authors analyzed the data. All authors drafted and revised the manuscript.

## References

1. Bech P. Fifty years with the Hamilton scales for anxiety and depression. A tribute to Max Hamilton. *Psychother Psychosom*. 2009 Jun;78(4):202-11.
2. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960 Feb;23(1):56-62.
3. Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol*. 1967 Dec;6(4):278-96.
4. Williams JBW. Standardizing the Hamilton Depression Rating Scale: past, present, and future. *Eur Arch Psychiatry Clin Neurosci*. 2001 Jun;251(Suppl. 2):6-12.
5. Zitman FG, Mennen MFG, Griez E, Hooijer C. The different versions of the Hamilton Depression Rating Scale. In: Bech P, Coppen A, editors. *The Hamilton Scales*. 1st ed. Berlin, Heidelberg: Springer-Verlag; 1990. pp. 28-34.
6. Rosenthal SH, Klerman GL. Endogenous features of depression in women. *Can Psychiatr Assoc J*. 1966 Dec;11(Suppl. 1):11-6.
7. Paykel ES, Klerman GL, Prusoff BA. Treatment setting and clinical depression. *Arch Gen Psychiatry*. 1970 Jan;22(1):11-21.
8. Paykel ES. The Clinical Interview for Depression: development, reliability and validity. *J Affect Disord*. 1985 Jul;9(1):85-96.
9. Guidi J, Fava GA, Bech P, Paykel ES. The Clinical Interview for Depression: A comprehensive review of studies and clinimetric properties. *Psychother Psychosom*. 2011;80(1):10-27.
10. Kovacs M, Rush AJ, Beck AT, Hollon SD. Depressed outpatients treated with cognitive therapy or pharmacotherapy: A one-year follow-up. *Arch Gen Psychiatry*. 1981 Jan;38(1):33-9.
11. Gelenberg AJ, Wojcik JD, Falk WE, Baldessarini RJ, Zeisel SH, Schoenfeld D, et al. Tyrosine for depression: a double-blind trial. *J Affect Disord*. 1990 Jun;19(2):125-32.
12. Thase ME, Frank E, Mallinger AG, Hamer T, Kupfer DJ. Treatment of imipramine-resistant recurrent depression, III: Efficacy of monoamine oxidase inhibitors. *J Clin Psychiatry*. 1992 Jan;53(1):5-11.
13. Bech P, Gram LF, Dein E, Jacobsen O, Vitger J, Bolwig TG. Quantitative rating of depressive states. Correlation between clinical assessment, Beck's self-rating scale and Hamilton's objective rating scale. *Acta Psychiatr Scand*. 1975 Mar;51(3):161-70.
14. Bech P, Allerup P, Gram LF, Reisby N, Rosenberg R, Jacobsen O, et al. The Hamilton Depression Scale. Evaluation of objectivity using logistic models. *Acta Psychiatr Scand*. 1981 Mar;63(3):290-9.
15. Bech P, Wilson P, Wessel T, Lunde M, Fava M. A validation analysis of two self-reported HAM-D<sub>6</sub> versions. *Acta Psychiatr Scand*. 2009 Apr;119(4):298-303.
16. Maier W, Philipp M. Improving the assessment of severity of depressive states: a reduction of the Hamilton Depression Scale. *Pharmacopsychiatry*. 1985;18(1):114-5.

17. Gibbons RD, Clark DC, Kupfer DJ. Exactly what does the Hamilton depression rating scale measure? *J Psychiatr Res.* 1993 Jul-Sep;27(3):259-73.
18. Bech P, Kastrup M, Rafaelsen OJ. Mini-compendium of rating scales for states of anxiety, depression, mania, schizophrenia, with corresponding DSM-III syndromes. *Acta Psychiatr Scand.* 1986;326(Suppl.):1-37.
19. Morriss R, Leese M, Chatwin J, Baldwin D, The THREAD Study Group. Inter-rater reliability of the Hamilton Depression Rating Scale as a diagnostic and outcome measure of depression in primary care. *J Affect Disord.* 2008 Dec;111(2-3):204-13.
20. Timmerby N, Andersen JH, Søndergaard S, Østergaard SD, Bech P. A systematic review of the clinimetric properties of the 6-item version of the Hamilton Depression Rating Scale (HAM-D<sub>6</sub>). *Psychother Psychosom.* 2017 May;86(3):141-9.
21. Miller IW, Bishop S, Norman WH, Maddever H. The Modified Hamilton Rating Scale for Depression: reliability and validity. *Psychiatry Res.* 1985 Feb;14(2):131-42.
22. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry.* 1988 Aug;45(8):742-7.
23. Whisman MA, Strosahl K, Fruzzetti AE, Schmaling KB, Jacobson NS, Miller DM. A structured interview version of the Hamilton Rating Scale for Depression: reliability and validity. *Psychol Assess.* 1989;1(3):238-41.
24. Potts MK, Daniels M, Burnam MA, Wells KB. A structured interview version of the Hamilton Depression Rating Scale: evidence of reliability and versatility of administration. *J Psychiatr Res.* 1990;24(4):335-50.
25. Williams JBW. Structured interview guides for the Hamilton Rating Scales. In: Bech P, Coppen A, editors. *The Hamilton Scales.* 1st ed. Berlin, Heidelberg: Springer-Verlag; 1990. pp. 48-63.
26. Williams JBW, Link MJ, Rosenthal NE, Amira L, Terman M. Structured Interview Guide for the Hamilton Depression Rating Scale-Seasonal Affective Disorder Version (SIGH-SAD). New York: New York State Psychiatric Institute; 1992.
27. Moberg PJ, Lazarus LW, Mesholam RI, Bilker W, Chuy IL, Neyman I, et al. Comparison of the standard and structured interview guide for the Hamilton Depression Rating Scale in depressed geriatric inpatients. *Am J Geriatr Psychiatry.* 2001;9(1):35-40.
28. Williams JBW, Kobak KA, Bech P, Engelhardt N, Evans K, Lipsitz J, et al. The GRID-HAMD: standardization of the Hamilton Depression Rating Scale. *Int Clin Psychopharmacol.* 2008 May;23(3):120-9.
29. Klerman GL, Weissman MM, Rounsaville BJ, Chevron ES. Interview format for the Hamilton Rating Scale for Depression. In: Klerman GL, Weissman MM, Rounsaville BJ, Chevron ES, editors. *Interpersonal Psychotherapy of Depression. A Brief, Focused, Specific Strategy.* Maryland: Rowman & Littlefield Publishers; 1984. pp. 223-233.



30. O'Hara MW, Rehm LP. Hamilton Rating Scale for Depression: reliability and validity of judgments of novice raters. *J Consult Clin Psychol.* 1983;51(2):318-9.
31. Robins LN, Helzer JE, Croughan J, Ratcliff KS. National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. *Arch Gen Psychiatry.* 1981 Apr;38(4):381-9.
32. Lambert MJ, Hatch DR, Kingston MD, Edwards BC. Zung, Beck, and Hamilton Rating Scales as measures of treatment outcome: A meta-analytic comparison. *J Consult Clin Psychol.* 1986;54(1):54-9.
33. Maier W. The Hamilton Depression Scale and its alternatives: a comparison of their reliability and validity. In: Bech P, Coppen A, editors. *The Hamilton Scales.* 1st ed. Berlin, Heidelberg: Springer-Verlag; 1990. pp. 64-71.
34. Grundy CT, Lunnen KM, Lambert MJ, Ashton JE, Tovey DR. The Hamilton Rating Scale for Depression: one scale or many? *Clin Psychol Sci Pract.* 1994 Dec;1(2):197-205.
35. Snaith RP. Present use of the Hamilton Depression Rating Scale: observations on method of assessment in research of depressive disorders. *Br J Psychiatry.* 1996 May;168(5):594-7.
36. Möller HJ. Methodological aspects in the assessment of severity of depression by the Hamilton Depression Scale. *Eur Arch Psychiatry Clin Neurosci* 2001 Jun;251(2):13-20.
37. Khan A, Khan SR, Shankles EB, Polissar NL. Relative sensitivity of the Montgomery-Åsberg Depression Rating Scale, the Hamilton Depression rating scale and the Clinical Global Impressions rating scale in antidepressant clinical trials. *Int Clin Psychopharmacol.* 2002 Nov;17(6):281-5.
38. Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *Am J Psychiatry.* 2004 Dec;161(12):2163-77.
39. Zimmerman M, Chelminski I, Posternak MA. A review of studies of the Hamilton Depression Rating Scale in healthy controls: implications for the definition of remission in treatment studies of depression. *J Nerv Ment Dis.* 2004 Sep;192(9):595-601.
40. Zimmerman M, Posternak MA, Chelminski I. Is it time to replace the Hamilton Depression Rating Scale as the primary outcome measure in treatment studies of depression? *J Clin Psychopharmacol.* 2005 Apr;25(2):105-10.
41. Bech P. Rating scales in depression: limitations and pitfalls. *Dialogues Clin Neurosci.* 2006 Jun;8(2), 207-15.
42. Shafer AB. Met-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *J Clin Psychol.* 2006 Jan;62(1):123-46.
43. López-Pina JA, Sánchez-Meca J, Rosa-Alcázar AI. The Hamilton Rating Scale for Depression: A meta-analytic reliability generalization study. *Int J Clin Health Psychol.* 2009;9(1):143-59.
44. Furukawa TA. Assessment of mood: guides for clinicians. *J Psychosom Res.* 2010 Jun;68(6):581-9.

45. Kriston L, von Wolff A. Not as golden as standards should be: interpretation of the Hamilton Rating Scale for Depression. *J Affect Disord.* 2011 Jan;128(1-2):175-7.
46. Trajković G, Starčević V, Latas M, Leštarević M, Ille T, Bukumirić Z, et al. Reliability of the Hamilton Rating Scale for Depression: a meta-analysis over a period of 49 years. *Psychiatry Res.* 2011 Aug;189(1):1-9.
47. Sharp R. The Hamilton rating scale for depression. *Occup Med.* 2015 Jun;65(4):340.
48. Vindbjerg E, Makransky G, Mortensen EL, Carlsson J. Cross-cultural psychometric properties of the Hamilton Depression Rating Scale. *Can J Psychiatry.* 2019;64(1):39-46.
49. Hedlund JL, Vieweg BW. The Hamilton rating scale for depression: a comprehensive review. *J Oper Psychiatry.* 1979;10(2):149-65.
50. Feinstein AR. T. Duckett Jones Memorial Lecture. The Jones criteria and the challenges of clinimetrics. *Circulation.* 1982 Jul;66(1):1-5.
51. Feinstein AR. An additional basic science for clinical medicine: IV. The development of clinimetrics. *Ann Intern Med.* 1983 Dec;99(6):843–8.
52. Feinstein AR. *Clinimetrics.* New Haven: Yale University Press; 1987.
53. Fava GA, Tomba E, Sonino N. Clinimetrics: the science of clinical measurements. *Int J Clin Pract.* 2012 Jan;66(1):11-5.
54. Fava GA, Ruini C, Rafanelli C. Psychometric theory is an obstacle to the progress of clinical research. *Psychother Psychosom.* 2004 May-Jun;73(3):145–8.
55. Fava GA, Carrozzino D, Lindberg L, Tomba E. The clinimetric approach to psychological assessment: A tribute to Per Bech, MD (1942-2018). *Psychother Psychosom.* 2018 Nov;87(6):321-6.
56. Bech, P. Modern psychometrics in clinimetrics: impact on clinical trials of antidepressants. *Psychother Psychosom.* 2004 May-Jun;73(3):134-8.
57. Fava GA, Belaise C. A discussion on the role of clinimetrics and the misleading effects of psychometric theory. *J Clin Epidemiol.* 2005 Aug;58(8):753-6.
58. Carrozzino D, Vassend O, Bjørndal F, Pignolo C, Olsen LR, Bech P. A clinimetric analysis of the Hopkins Symptom Checklist (SCL-90-R) in general population studies (Denmark, Norway, and Italy). *Nord J Psychiatry* 2016;70(5):374-9.
59. Fava GA, Carrozzino D, Lindberg L, Tomba E. Reply to the Letter to the Editor: “Is a single-item measure of self-rated mental health useful from a clinimetric perspective?” *Psychother Psychosom.* 2019 Jun;88(3):179
60. Fleck MP, Carrozzino D, Fava GA. The challenge of measurement in psychiatry: the lifetime accomplishments of Per Bech (1942-2018). *Braz J Psychiatry.* 2019 Sep-Oct;41(5):369-72.

61. Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. *J Clin Epidemiol*. 1992 Nov;45(11):1201-18.
62. Kellner R, Sheffield BF. A self-rating scale of distress. *Psychol Med*. 1973 Feb;3(1):88-100.
63. Fava GA, Guidi J, Rafanelli C, Rickels K. The clinical inadequacy of the placebo model and the development of an alternative conceptual framework. *Psychother Psychosom*. 2017 Nov;86(6):332-40.
64. Fava GA, Tomba E, Bech P. Clinical pharmacopsychology: conceptual foundations and emerging tasks. *Psychother Psychosom*. 2017 May;86(3):134-40.
65. Guidi J, Brakemeier EL, Bockting CL, Cosci F, Cuijpers P, Jarrett RB, et al. Methodological recommendations for trials of psychological interventions. *Psychother Psychosom*. 2018 Sep;87(5):276-84.
66. Prasad MK, Udupa K, Kishore KR, Thirthalli J, Sathyaprabha TN, Gangadhar BN. Inter-rater reliability of Hamilton depression rating scale using video-recorded interviews—Focus on rater-blinding. *Indian J Psychiatry*. 2009 Jul-Sep;51(3):191-4.
67. Knesevich JW, Biggs JT, Clayton PJ, Ziegler VE. Validity of the Hamilton Rating Scale for depression. *Br J Psychiatry*. 1977 Jul;131(1):49-52.
68. Ziegler VE, Meyer DA, Rosen SH, Biggs JT. Reliability of video taped Hamilton ratings. *Biol Psychiatry*. 1978;13(1):119-22.
69. Bech P, Bolwig TG, Kramp P, Rafaelsen OJ. The Bech-Rafaelsen Mania Scale and the Hamilton Depression Scale: evaluation of homogeneity and inter-observer reliability. *Acta Psychiatr Scand*. 1979 Apr;59(4):420-30.
70. Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, et al. Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res*. 1982;17(1):37-49.
71. Cicchetti DV, Prusoff BA. Reliability of depression and associated clinical symptoms. *Arch Gen Psychiatry*. 1983 Sep;40(9):987-90.
72. Craig TJ, Richardson MA, Pass R, Bregman Z. Measurement of mood and affect in schizophrenic inpatients. *Am J Psychiatry*. 1985;142(11):1272-77.
73. Rehm LP, O'Hara MW. Item characteristics of the Hamilton rating scale for depression. *J Psychiatr Res*. 1985;19(1):31-41.
74. Deluty BM, Deluty RH, Carver CS. Concordance between clinicians' and patients' ratings of anxiety and depression as mediated by private self-consciousness. *J Pers Assess*. 1986;50(1):93-106.
75. Gottlieb GL, Gur RE, Gur RC. Reliability of psychiatric scales in patients with dementia of the Alzheimer type. *Am J Psychiatry*. 1988;145(7):857-60.
76. Ramos-Brieva JA, Cordero-Villafafila A. A new validation of the Hamilton Rating Scale for Depression. *J Psychiatr Res*. 1988;22(1):21-8.

77. Zheng Y, Zhao J, Phillips M, Liu J, Cai M, Sun S, et al. Validity and reliability of the Chinese Hamilton depression rating scale. *Br J Psychiatry*. 1988 May;152(5):660-4.
78. Leung CM, Wing YK, Kwong PK, Shum ALK. Validation of the Chinese-Cantonese version of the Hospital Anxiety and Depression Scale and comparison with the Hamilton Rating Scale of Depression. *Acta Psychiatr Scand*. 1999 Dec;100(6):456-61.
79. Baca-García E, Blanco C, Sáiz-Ruiz J, Rico F, Diaz-Sastre C, Cicchetti DV. Assessment of reliability in the clinical evaluation of depressive symptoms among multiple investigators in a multicenter clinical trial. *Psychiatry Res*. 2001 Jun;102(2):163-73.
80. Pancheri P, Picardi A, Pasquini M, Gaetano P, Biondi M. Psychopathological dimensions of depression: a factor study of the 17-item Hamilton depression rating scale in unipolar depressed outpatients. *J Affect Disord*. 2002 Feb;68(1):41-7.
81. Wagner S, Helmreich I, Lieb K, Tadic A. Standardized Rater Training for the Hamilton Depression Rating Scale (HAMD<sub>17</sub>) and the Inventory of Depressive Symptoms (IDSC<sub>30</sub>). *Psychopathology*. 2011;44(1):68-70.
82. Tabuse H, Kalali A, Azuma H, Ozaki N, Iwata N, Naitoh H, et al. The new GRID Hamilton Rating Scale for Depression demonstrates excellent inter-rater reliability for inexperienced and experienced raters before and after training. *Psychiatry Res*. 2007 Sep;153(1):61-7.
83. Kørner A, Nielsen BM, Eschen F, Møller-Madsen S, Stender A, Christensen EM, et al. Quantifying depressive symptomatology: inter-rater reliability and inter-item correlations. *J Affect Disord*. 1990 Oct;20(2):143-9.
84. Koenig HG, Pappas P, Holsinger T, Bachar JR. Assessing diagnostic approaches to depression in medically ill older adults: how reliably can mental health professionals make judgments about the cause of symptoms? *J Am Geriatr Soc*. 1995 May;43(5):472-8.
85. Fuglum E, Rosenberg C, Damsbo N, Stage K, Lauritzen L, Bech P, et al. Screening and treating depressed patients. A comparison of two controlled citalopram trials across treatment settings: hospitalized patients vs. patients treated by their family doctors. *Acta Psychiatr Scand*. 1996 Jul;94(1):18-25.
86. Bent-Hansen J, Bech P. Validity of the definite and semidefinite questionnaire version of the Hamilton Depression Scale, the Hamilton Subscale and the Melancholia Scale. Part I. *Eur Arch Psychiatry Clin Neurosci*. 2011;261(1):37-46.
87. Chipman A, Paykel ES. How ill is the patient at this time? Cues determining clinician's global judgments. *J Consult Clin Psychol*. 1974;42(5):669-74.
88. Paykel ES, Mangen SP, Griffith JH, Burns TP. Community psychiatric nursing for neurotic patients: a controlled trial. *Br J Psychiatry*. 1982;140(6):573-81.
89. Akdemir A, Türkçapar MH, Örsel SD, Demirergi N, Dag I, Özbay MH. Reliability and validity of the Turkish version of the Hamilton Depression Rating Scale. *Compr Psychiatry*. 2001 Mar;42(2):161-5.

90. Rohan KJ, Rough JN, Evans M, Ho SY, Meyerhoff J, Roberts LM, et al. A protocol for the Hamilton Rating Scale for Depression: Item scoring rules, Rater training, and outcome accuracy with data on its application in a clinical trial. *J Affect Disord.* 2016 Aug;200:111-8.
91. Bech P. *Clinical Psychometrics.* Oxford: Wiley-Blackwell; 2012.
92. van Agt HM, Essink-Bot ML, Krabbe PF, Bonsel GJ. Test-retest reliability of health state valuations collected with the EuroQol questionnaire. *Soc Sci Med.* 1994;39(11):1537-44.
93. Shankman SA, Klein DN. The impact of comorbid anxiety disorders on the course of dysthymic disorder: a 5-year prospective longitudinal study. *J Affect Disord.* 2002 Jul;70(2):211-7.
94. Tomba E, Bech P. Clinimetrics and clinical psychometrics: macro- and micro-analysis. *Psychother Psychosom.* 2012 Oct;81(6):333-43.
95. Bech, P. The use of rating scales in affective disorders. *Eur Psychiatr Rev.* 2008;1:14-8.
96. Licht RW, Qvitzau S, Allerup P, Bech P. Validation of the Bech-Rafaelsen Melancholia Scale and the Hamilton Depression Scale in patients with major depression; is the total score a valid measure of illness severity? *Acta Psychiatr Scand.* 2005 Feb;111(2):144-9.
97. Ruhé HG, Dekker JJ, Peen J, Holman R, de Jonghe F. Clinical use of the Hamilton Depression Rating Scale: is increased efficiency possible? A post hoc comparison of Hamilton Depression Rating Scale, Maier and Bech subscales, Clinical Global Impression, and Symptom Checklist-90 scores. *Compr Psychiatry.* 2005 Nov-Dec;46(6):417-27.
98. Bobes J, Bulbena A, Luque A, Dal-Re R, Ballesteros J, Ibarra N, et al. The sufficiency of the HAM-D<sub>6</sub> as an outcome instrument in the acute therapy of antidepressants in the outpatient setting. *Int J Psychiatry Clin Pract.* 2007;11(2):146-50.
99. Bobo WV, Angleró GC, Jenkins G, Hall-Flavin DK, Weinshilboum R, Biernacka JM. Validation of the 17-item Hamilton Depression Rating Scale definition of response for adults with major depressive disorder using equipercentile linking to clinical global impression scale ratings: analysis of pharmacogenomic research network antidepressant medication pharmacogenomic study (PGRN-AMPS) data. *Hum Psychopharm Clin.* 2016 May;31(3):185-92.
100. Kyle PR, Lemming OM, Timmerby N, Søndergaard S, Andreasson K, Bech P. The validity of the different versions of the Hamilton Depression Scale in separating remission rates of placebo and antidepressants in clinical trials of major depression. *J Clin Psychopharmacol.* 2016 Oct;36(5):453-6.
101. Carroll BJ, Fielding JM, Blashki TG. Depression rating scales: a critical review. *Arch Gen Psychiatry.* 1973 Mar;28(3):361-6.
102. Thase ME, Hersen M, Bellack AS, Himmelhoch JM, Kupfer DJ. Validation of a Hamilton subscale for endogenomorphic depression. *J Affect Disord.* 1983 Aug;5(3):267-78.
103. Endicott J, Spitzer RL, Fleiss JL, Cohen J. The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disturbance. *Arch Gen Psychiatry.* 1976 Jun;33(6):766-71.

104. Ganchrow JR, Steiner JE, Kleiner M, Edelstein EL. A multidisciplinary approach to the expression of pain in psychic depression. *Percept Mot Ski*. 1978 Dec;47(2):379-90.
105. Fava GA, Kellner R, Munari F, Pavan L. The Hamilton Depression Rating Scale in normals and depressives. *Acta Psychiatr Scand*. 1982 Jul;66(1):26-32.
106. Rush AJ, Giles DE, Schlessler MA, Fulton CL, Weissenburger J, Burns C. The inventory for depressive symptomatology (IDS): preliminary findings. *Psychiatry Res*. 1986 May;18(1): 65-87.
107. Carneiro AM, Fernandes F, Moreno RA. Hamilton Depression Rating Scale and Montgomery–Asberg depression rating scale in depressed and bipolar I patients: psychometric properties in a Brazilian sample. *Health Qual Life Outcomes* 2015 Apr;13(1):42.
108. Carrozzino D. Clinimetric approach to rating scales for the assessment of apathy in Parkinson's disease: A systematic review. *Prog Neuro-Psychopharmacol Biol Psychiatry*. 2019 Aug;94:109641.
109. Zimmerman M, Martinez JH, Young D, Chelminski I, Dalrymple K. Severity classification on the Hamilton depression rating scale. *J Affect Disord*. 2013 Sep;150(2):384-8.
110. Freeling P, Rao BM, Paykel ES, Sireling LI, Burton RH. Unrecognised depression in general practice. *BMJ*. 1985 Jun;290:1880–3.
111. Freeling P. Diagnosis and treatment of depression in general-practice. *Br J Psychiatry*. 1993 Jul;163(Suppl. 20):14–9.
112. Sireling LI, Paykel ES, Freeling P, Rao BM, Patel SP. Depression in general practice: case thresholds and diagnosis. *Br J Psychiatry*. 1985 Aug;147(2):113–9.
113. Tylee AT, Freeling P, Kerry S. Why do general practitioners recognize major depression in one woman patient yet miss it in another? *Br J Gen Pract*. 1993 Aug;43(373):327–30.
114. Brophy JJ. Personality disorder, symptoms and dexamethasone suppression in depression. *J Affect Disord*. 1994 May;31(1):19–27.
115. Michael A, Jenaway A, Paykel ES, Herbert J. Altered salivary dehydroepiandrosterone levels in major depression in adults. *Biol Psychiatry*. 2000 Nov;48(10):989–95.
116. Evers AW, Colloca L, Blease C, Annoni M, Atlas LY, Benedetti F, et al. Implications of placebo and nocebo effects for clinical practice: expert consensus. *Psychother Psychosom*. 2018 Aug;87(4):204-10.
117. Trivedi MH, South C, Jha MK, Rush AJ, Cao J, Kurian B, et al. A novel strategy to identify placebo responders: prediction index of clinical and biological markers in the EMBARC trial. *Psychother Psychosom*. 2018 Sep;87(5):285-95.
118. Montgomery SA, Åsberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry*. 1979 Apr;134(4):382-9.

119. van Moffaert M, de Wilde J, Vereecken A, Dierick M, Evrard JL, Wilmotte J, et al. Mirtazapine is more effective than trazodone: a double-blind controlled study in hospitalized patients with major depression. *Int Clin Psychopharmacol*. 1995 Mar;10(1):3-9.
120. Feiger AD. A double-blind comparison of gepirone extended release, imipramine, and placebo in the treatment of outpatient major depression. *Psychopharmacol Bull*. 1996 Jan;32(4):659-65.
121. Wilcox CS, Ferguson JM, Dale JL, Heiser JF. A double-blind trial of low-and high-dose ranges of gepirone-ER compared with placebo in the treatment of depressed outpatients. *Psychopharmacol Bull*. 1996 Jan;32(3):335-42.
122. Lecrubier Y, Clerc G, Didi R, Kieser M. Efficacy of St. John's wort extract WS 5570 in major depression: a double-blind, placebo-controlled trial. *Am J Psychiatry*. 2002 Aug;159(8):1361-6.
123. Feiger AD, Heiser JF, Shrivastava RK, Weiss KJ, Smith WT, Sitsen JM, et al. Gepirone extended-release: new evidence for efficacy in the treatment of major depressive disorder. *J Clin Psychiatry*. 2003 Mar;64(3):243-9.
124. Liebowitz MR, Yeung PP, Entsuah R. A randomized, double-blind, placebo-controlled trial of desvenlafaxine succinate in adult outpatients with major depressive disorder. *J Clin Psychiatry*. 2007 Nov;68(11):1663-72.
125. Liebowitz MR, Manley AL, Padmanabhan SK, Ganguly R, Tummala R, Tourian KA. Efficacy, safety, and tolerability of desvenlafaxine 50 mg/day and 100 mg/day in outpatients with major depressive disorder. *Curr Med Res Opin*. 2008 May;24(7):1877-90.
126. Fabre LF, Abuzzahab FS, Amin M, Claghorn JL, Mendels J, Petrie WM, et al. Sertraline safety and efficacy in major depression: a double-blind fixed-dose comparison with placebo. *Biol Psychiatry*. 1995 Nov;38(9):592-602.
127. Entsuah R, Shaffer M, Zhang J. A critical examination of the sensitivity of unidimensional subscales derived from the Hamilton Depression Rating Scale to antidepressant drug effects. *J Psychiatr Res*. 2002 Nov-Dec;36(6):437-48.
128. Santen G, Gomeni R, Danhof M, Della Pasqua O. Sensitivity of the individual items of the Hamilton Depression Rating Scale to response and its consequences for the assessment of efficacy. *J Psychiatr Res*. 2008 Oct;42(12):1000-9.
129. Sheehan DV, Rozova A, Gossen ER, Gibertini M. The efficacy and tolerability of once-daily controlled-release trazodone for depressed mood, anxiety, insomnia, and suicidality in major depressive disorder. *Psychopharmacol Bull*. 2009 Jan;42(4):5-22.
130. Inamdar A, Merlo-Pich E, Gee M, Makumi C, Mistry P, Robertson J, et al. Evaluation of antidepressant properties of the p38 MAP kinase inhibitor losmapimod (GW856553) in major depressive disorder: Results from two randomised, placebo-controlled, double-blind, multicentre studies using a Bayesian approach. *J Psychopharmacol*. 2014 Jun;28(6):570-81.

131. Luckenbaugh DA, Ameli R, Brutsche NE, Zarate CA. Rating depression over brief time intervals with the Hamilton Depression Rating Scale: Standard vs. abbreviated scales. *J Psychiatr Res.* 2015 Feb;61:40-5.
132. Preskorn S, Macaluso M, Mehra DV, Zammit G, Moskal JR, Burch RM. Randomized proof of concept trial of GLYX-13, an N-methyl-D-aspartate receptor glycine site partial agonist, in major depressive disorder nonresponsive to a previous antidepressant agent. *J Psychiatr Pract.* 2015;21(2):140-9.
133. Helmreich I, Wagner S, Mergl R, Allgaier AK, Hautzinger M, Henkel V, et al. The Inventory of Depressive Symptomatology (IDS-C<sub>28</sub>) is more sensitive to changes in depressive symptomatology than the Hamilton Depression Rating Scale (HAMD<sub>17</sub>) in patients with mild major, minor or subsyndromal depression. *Eur Arch Psychiatry Clin Neurosci.* 2011;261(5):357-67.
134. Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The inventory of depressive symptomatology (IDS): psychometric properties. *Psychol Med.* 1996 May;26(3):477-86.
135. Liu J, Xiang YT, Wang G, Zhu XZ, Ungvari GS, Kilbourne AM, et al. Psychometric properties of the Chinese versions of the Quick Inventory of Depressive Symptomatology–Clinician Rating (C-QIDS-C) and Self-Report (C-QIDS-SR). *J Affect Disord.* 2013 May;147(1-3):421-4.
136. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry.* 2003 Sep;54(5):573-83.
137. Trivedi MH, Rush AJ, Ibrahim HM, Carmody TJ, Biggs MM, Suppes T, et al. The Inventory of Depressive Symptomatology, clinician Rating (IDS-C) and self-report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, clinician rating (QIDS-C) and self-report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. *Psychol Med.* 2004 Jan;34(1):73-82.
138. Østergaard SD, Bech P, Trivedi MH, Wisniewski SR, Rush AJ, Fava M. Brief, unidimensional melancholia rating scales are highly sensitive to the effect of citalopram and may have biological validity: implications for the research domain criteria (RDoC). *J Affect Disord.* 2014 Jul;163:18-24.
139. Bech P, Fava M, Trivedi MH, Wisniewski SR, Rush AJ. Outcomes on the pharmacopsychometric triangle in bupropion-SR vs. bupropion augmentation of citalopram in the STAR\* D trial. *Acta Psychiatr Scand.* 2012 Apr;125(4):342-8.
140. Rickels K, Chung HR, Csanalosi IB, Hurowitz AM, London J, Wiseman K, et al. Alprazolam, diazepam, imipramine, and placebo in outpatients with major depression. *Arch Gen Psychiatry.* 1987 Oct;44(10):862-6.
141. Bech P, Cialdella P, Haugh MC, Birkett MA, Hours A, Boissel JP, et al. Meta-analysis of randomised controlled trials of fluoxetine v. placebo and tricyclic antidepressants in the short-term treatment of major depression. *Br J Psychiatry.* 2000 May;176(5):421-8.



142. Faries D, Herrera J, Rayamajhi J, DeBrotta D, Demitrack M, Potter WZ. The responsiveness of the Hamilton Depression Rating Scale. *J Psychiatr Res.* 2000 Jan;34(1):3-10.
143. Schneider LS, Nelson JC, Clary CM, Newhouse P, Krishnan KRR, Shiovitz T, et al. An 8-week multicenter, parallel-group, double-blind, placebo-controlled study of sertraline in elderly outpatients with major depression. *Am J Psychiatry.* 2003 Jul;160(7):1277-85.
144. Bech P, Kajdasz DK, Porsdal V. Dose-response relationship of duloxetine in placebo-controlled clinical trials in patients with major depressive disorder. *Psychopharmacol.* 2006 Sep; 188(3):273-80.
145. Bech P, Boyer P, Germain JM, Padmanabhan K, Haudiquet V, Pitrosky B, et al. HAM-D<sub>17</sub> and HAM-D<sub>6</sub> sensitivity to change in relation to desvenlafaxine dose and baseline depression severity in major depressive disorder. *Pharmacopsychiatry.* 2010;43(7):271-6.
146. Thase ME, Ninan PT, Musgnung JJ, Trivedi MH. Remission with venlafaxine extended release or selective serotonin reuptake inhibitors in depressed patients: a randomized, open-label study. *Prim Care Companion CNS Disord.* 2011;13(1):PCC.10m00979.
147. Chouinard G, Saxena BM, Nair NPV, Kutcher SP, Bakish D, Bradwejn J, et al. A Canadian multicentre placebo-controlled study of a fixed dose of brofaromine, a reversible selective MAO-A inhibitor, in the treatment of major depression. *J Affect Disord.* 1994 Oct;32(2):105-14.
148. Feiger AD, Rickels K, Rynn MA, Zimbhoff DL, Robinson DS. Selegiline transdermal system for the treatment of major depressive disorder: an 8-week, double-blind, placebo-controlled, flexible-dose titration trial. *J Clin Psychiatry.* 2006 Sep;67(9):1354-61.
149. Søndergaard MP, Jarden JO, Martiny K, Andersen G, Bech P. Dose response to adjunctive light therapy in citalopram-treated patients with post-stroke depression. *Psychother Psychosom.* 2006 Jun;75(4):244-8.
150. Edwards BC, Lambert MJ, Moran PW, McCully T, Smith KC, Ellingson AG. A meta-analytic comparison of the Beck Depression Inventory and the Hamilton Rating Scale for Depression as measures of treatment outcome. *Br J Clin Psychol.* 1984 May;23(2):93-9.
151. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. Beck Depression Inventory (BDI). *Arch Gen Psychiatry.* 1961;4(6):561-71.
152. Sayer NA, Sackeim HA, Moeller JR, Prudic J, Devanand DP, Coleman EA, et al. The relations between observer-rating and self-report of depressive symptomatology. *Psychol Assess.* 1993;5(3):350-60.
153. Hollyman JA, Freeling P, Paykel ES, Bhat A, Sedgwick P. Double-blind placebo-controlled trial of amitriptyline among depressed patients in general-practice. *J R Coll Gen Pract.* 1988 Sep;38(314):393-7.
154. Paykel ES, Hollyman JA, Freeling P, Sedgwick P. Predictors of therapeutic benefit from amitriptyline in mild depression: a general practice placebo-controlled trial. *J Affect Disord.* 1988 Jan-Feb;14(1):83-95.

155. Paykel ES. Use of the Hamilton Depression Scale in general practice. In: Bech P, Coppen A, editors. *The Hamilton Scales*. 1st ed. Berlin, Heidelberg: Springer-Verlag; 1990. pp. 4-47.
156. Paykel ES, Scott J, Teasdale JD, Johnson AL, Garland A, Moore R, et al. Prevention of relapse in residual depression by cognitive therapy: a controlled trial. *Arch Gen Psychiatry*. 1999 Sep;56(9):829-35.
157. Scott J, Teasdale JD, Paykel ES, Johnson AL, Abbott R, Hayhurst H, et al. Effects of cognitive therapy on psychological symptoms and social functioning in residual depression. *Br J Psychiatry*. 2000 Nov;177(5):440-6.
158. Paykel ES, Scott J, Cornwall PL, Abbott R, Crane C, Pope M, et al. Duration of relapse prevention after cognitive therapy in residual depression: follow-up of controlled trial. *Psychol Med*. 2005 Jan;35(1):59-68.
159. Carrozzino D, Siri C, Bech P. The prevalence of psychological distress in Parkinson's disease patients: the brief symptom inventory (BSI-18) versus the Hopkins Symptom Checklist (SCL-90-R). *Prog. Neuropsychopharmacol Biol Psychiatry*. 2019 Jan;88:96-101.
160. Rasch G. *Probabilistic models for some intelligence and attainment tests*. Expanded edition. Chicago: The University of Chicago Press; 1980.
161. Mokken RJ. *A theory and procedure of scale analysis*. Berlin: Mouton; 1971.
162. Rocha NS, Power MJ, Bushnell DM, Fleck MP. Cross-cultural evaluation of the WHOQOL-BREF domains in primary care depressed patients using Rasch analysis. *Med Decis Making*. 2012 Jan-Feb;32(1):41-55.
163. Carrozzino D, Svicher A, Patierno C, Berrocal C, Cosci F. The Euthymia Scale: A Clinimetric Analysis. *Psychother Psychosom*. 2019 Apr;88(2):119-21.
164. Loevinger J. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychol Bull*. 1948 Nov;45(6):507-29.
165. Bech P, Allerup P, Reisby N, Gram LF. Assessment of symptom change from improvement curves on the Hamilton depression scale in trials with antidepressants. *Psychopharmacology*. 1984 Oct;84(2):276-81.
166. Bech P, Tanghøj P, Andersen H, Overø K. Citalopram dose-response revisited using an alternative psychometric approach to evaluate clinical effects of four fixed citalopram doses compared to placebo in patients with major depression. *Psychopharmacology*. 2002 Jul;163(1):20-5.
167. Bech P, Tanghøj P, Cialdella P, Andersen HF, Pedersen AG. Escitalopram dose-response revisited: an alternative psychometric approach to evaluate clinical effects of escitalopram compared to citalopram and placebo in patients with major depression. *Int J Neuropsychopharmacol*. 2004 Sep;7(3):283-90.
168. Bech P, Fava M, Trivedi MH, Wisniewski SR, Rush AJ. Factor structure and dimensionality of the two depression scales in STAR\*D using level 1 datasets. *J Affect Disord*. 2011 Aug;132(3):396-400.

169. Martiny K, Refsgaard E, Lund V, Lunde M, Sørensen L, Thougard B, et al. The day-to-day acute effect of wake therapy in patients with major depression using the HAM-D<sub>6</sub> as primary outcome measure: results from a randomised controlled trial. *PloS One*. 2013 Jun;8(6):e67264.
170. Bech P, Allerup P, Larsen ER, Csillag C, Licht RW. The Hamilton Depression Scale (HAM-D) and the Montgomery-Åsberg Depression Scale (MADRS). A psychometric re-analysis of the european genome-based therapeutic drugs for depression study using rasch analysis. *Psychiatry Res*. 2014 Jul;217(3):226-32.
171. Østergaard SD, Bech P, Miskowiak KW. Fewer study participants needed to demonstrate superior antidepressant efficacy when using the Hamilton melancholia subscale (HAM-D<sub>6</sub>) as outcome measure. *J Affect Disord*. 2016 Jan;190:842-5.
172. Holmskov J, Licht RW, Andersen K, Stage TB, Nilsson FM, Stage KB, et al. Diagnostic conversion to bipolar disorder in unipolar depressed patients participating in trials on antidepressants. *Eur Psychiatry*. 2017 Feb;40:76-81.
173. da Silva AK, Reche M, da Silva Lima AF, Fleck MP, Capp E, Shansis FM. Assessment of the psychometric properties of the 17-and 6-item Hamilton Depression Rating Scales in major depressive disorder, bipolar depression and bipolar depression with mixed features. *J Psychiatr Res*. 2019 Jan;108:84-9.
174. Bech P, Paykel E, Sireling L, Yiend J. Rating scales in general practice depression: Psychometric analyses of the Clinical Interview for Depression and the Hamilton Rating Scale. *J Affect Disord*. 2015;171:68-73.
175. Prusoff BA, Klerman GL, Paykel ES. Concordance between clinical assessments and patients' self-report in depression. *Arch Gen Psychiatry*. 1972 Jun;26(6):546-52.
176. Kellner R. A Symptom Questionnaire. *J Clin Psychiatry*. 1987 Jul;48(7):268-74.
177. Benasi G, Fava GA, Rafanelli C. Kellner's Symptom Questionnaire, a highly sensitive patient-reported outcome measure: systematic review of clinimetric properties. *Psychother Psychosom*. 2020 Feb; doi: 10.1159/000506110.
178. Zung WWK. A self-rating depression scale. *Arch Gen Psychiatry*. 1965 Jan;12(1):63-70.
179. Carroll BJ, Feinberg M, Smouse PE, Rawson SG, Greden JF. The Carroll rating scale for depression I. Development, reliability and validation. *Br J Psychiatry*. 1981 Mar;138(3):194-200.
180. Berard RMF, Ahmed N. Hospital Anxiety and Depression Scale (HADS) as a screening instrument in a depressed adolescent and young adult population. *Int J Adolesc Med Health*. 1995;8(3):157-66.
181. Brown C, Schulberg HC, Madonia MJ. Assessment depression in primary care practice with the Beck Depression Inventory and the Hamilton Rating Scale for Depression. *Psychol Assess*. 1995;7(1):59-65.
182. Hotopf M, Sharp D, Lewis G. What's in a name? A comparison of four psychiatric assessments. *Soc Psychiatry Psychiatr Epidemiol*. 1997 Dec;33(1):27-31.

183. Kobak KA, Greist JH, Jefferson JW, Mundt JC, Katzelnick DJ. Computerized assessment of depression and anxiety over the telephone using interactive voice response. *MD Comput.* 1999 May;16(3):64-8.
184. Paykel ES, Penrose RJ, Rassaby E. Depressive classification and prediction of response to phenelzine. *Br J Psychiatry.* 1979 Jun;134(6):572-81.
185. Rowan PR, Paykel ES, Parker RR. Phenelzine and amitriptyline: effects on symptoms of neurotic depression. *Br J Psychiatry.* 1982 May;140(5):475-83.
186. Norton KRW, Sireling LI, Bhat AV, Rao B, Paykel ES. A double-blind comparison of fluvoxamine, imipramine and placebo in depressed patients. *J Affect Disord.* 1984 Dec;7(3-4):297-308.
187. Prusoff BA, Weissman MM. Pharmacological treatment of anxiety in depressed outpatients. In: Klein DF, Rabkin J, editors. *Anxiety - New Research and Changing Concepts.* New York: Raven Press; 1981. pp. 341-53.
188. Nierenberg AA, Sonino N. From clinical observations to clinimetrics: a tribute to Alvan R. Feinstein, MD. *Psychother Psychosom.* 2004;73(3):131-3.
189. Fava GA, Rafanelli C, Tomba E. The clinical process in psychiatry: a clinimetric approach. *J Clin Psychiatry.* 2012 Feb;73(2):177-84.
190. Bech P. *Rating Scales for Psychopathology, Health Status and Quality of Life.* Berlin: Springer; 1993.
191. Bech P. *Measurement-based care in mental disorders.* New York: Springer-Verlag; 2016.
192. Corruble E, Hardy P. Why the Hamilton depression rating scale endures. *Am J Psychiatry.* 2005 Dec;162(12):2394.
193. Carroll B. Why the Hamilton depression rating scale endures. *Am J Psychiatry.* 2005 Dec;162(12):2395-6.
194. Concato J, Horwitz RI. Limited usefulness of meta-analysis for informing patient care. *Psychother Psychosom.* 2019 Sep;88(5):257-62.
195. Lobitz G, Armstrong K, Concato J, Singer BH, Horwitz RI. The biological and biographical basis of precision medicine. *Psychother Psychosom.* 2019 Nov;88(6):333-40.
196. Bekhuis E, Schoevers R, de Boer M, Peen J, Dekker J, Van H, et al. Symptom-specific effects of psychotherapy versus combined therapy in the treatment of mild to moderate depression: a network approach. *Psychother Psychosom.* 2018 Mar;87(2):121-3.
197. Contreras A, Nieto I, Valiente C, Espinosa R, Vazquez C. The study of psychopathology from the network analysis perspective: a systematic review. *Psychother Psychosom.* 2019 Apr;88(2):71-83.

198. Guidi J, Tomba E, Fava GA. The sequential integration of pharmacotherapy and psychotherapy in the treatment of major depressive disorder: a meta-analysis of the sequential model and a critical review of the literature. *Am J Psychiatry*. 2016 Feb;173(2):128-37.
199. Fava GA, Bech P. The concept of euthymia. *Psychother Psychosom*. 2016 Jan;85(1):1-5.
200. Bech P, Carrozzino D, Austin SF, Møller SB, Vassend O. Measuring euthymia within the neuroticism scale from the NEO Personality Inventory: a mokken analysis of the norwegian general population study for scalability. *J Affect Disord*. 2016 Mar;193:99-102.
201. Fava GA, Guidi J. The pursuit of euthymia. *World Psychiatry*. 2020 Feb;19(1):40-50.
202. Guidi J, Piolanti A, Gostoli S, Schamong I, Brakemeier EL. Mental pain and euthymia as transdiagnostic clinimetric indices in primary care. *Psychother Psychosom*. 2019 Aug;88(4):252-3.
203. Fava GA, Tomba E, Brakemeier EL, Carrozzino D, Cosci F, Eöry A, et al. Mental Pain as a Transdiagnostic Patient-Reported Outcome Measure. *Psychother Psychosom*. 2019 Nov;88(6):341-9.

**Table 1.** Items of the most widely used HAM-D versions

		Hamilton [2]	Hamilton [2, 3]	Bech et al. [13, 14]	Miller et al. [21]	Bech et al. [18]	Williams [22]	Potts et al. [24]	Gelenberg [11]	Williams et al. [26]	Moberg et al. [27]	Morriss et al. [19]	Timmerby et al. [20]
		HAM-D <sub>21</sub>	HAM-D <sub>17</sub>	HAM-D <sub>6</sub>	MHRSD <sub>25</sub>	HAM-D <sub>23</sub>	SIGH-D <sub>21</sub>	SI-HDRS <sub>17</sub>	HAM-D <sub>27</sub>	SIGH-SAD <sub>29</sub>	HAM-D <sub>24</sub>	HDRS <sub>17</sub>	HAM-D <sub>6</sub>
		Unstructured version with ratings only	Unstructured version with ratings only	Unstructured version with ratings only	Structured version with structured interview	Structured version with anchor points and item definitions	Structured version with structured interview	Structured version with structured interview	Unstructured version with ratings only	Structured version with structured interview	Structured version with structured interview	Structured version with semi-structured interview and anchor points	Structured version with semi-structured interview and anchor points
N°	Items	Range of scores											
1	Depressed mood	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4
2	Feelings of guilt	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-3	0-4	0-4	0-4
3	Suicide	0-4	0-4	-	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	-
4	Insomnia, early	0-2	0-2	-	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	-
5	Insomnia, middle	0-2	0-2	-	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	-
6	Insomnia, late	0-2	0-2	-	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	-
7	Work and interests	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4
8	Psychomotor retardation	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4
9	Psychomotor agitation	0-2	0-2	-	0-2	0-4	0-4	0-4	0-4	0-4	0-4	0-4	-
10	Anxiety, psychic	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4
11	Anxiety, somatic	0-4	0-4	-	0-4	0-4	0-4	-	0-4	0-4	0-4	0-4	-
12	Somatic symptoms, GI	0-2	0-2	-	-	0-2	0-2	-	0-2	0-2	0-2	0-4	-
13	Somatic symptoms, general	0-2	0-2	-	-	0-2	0-2	0-4	0-2	0-2	0-2	0-2	-
14	Genital symptoms	0-2	0-2	-	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	-
15	Hypochondriasis	0-4	0-4	-	0-4	0-4	0-4	0-4	0-4	0-4	0-4	0-4	-
16	Loss of weight	0-2	0-2	-	0-2	0-2	0-3	-	0-2	0-2	0-3	0-2	-
17	Insight	0-2	0-2	-	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-2	-
18	Diurnal variation	0-2	-	-	0-2	-	0-2	-	0-2	-	0-2	-	-
19	Depersonalization/derealization	0-4	-	-	-	-	0-4	-	0-4	0-4	0-4	-	-
20	Paranoid symptoms	0-4	-	-	-	-	0-3	-	0-3	0-3	0-3	-	-
21	Obsessive/compulsive	0-2	-	-	-	-	0-2	-	0-2	0-2	0-2	-	-
22	Tiredness and pains	-	-	0-2	0-2	0-4	-	0-2	-	-	-	-	0-2
23	Distinct quality of mood	-	-	-	0-2	-	-	-	-	-	-	-	-
24	Lack of reactivity	-	-	-	0-2	-	-	-	-	-	-	-	-
25	Worthlessness	-	-	-	0-4	-	-	-	-	-	0-4	-	-
26	Helplessness	-	-	-	0-4	-	-	-	-	-	0-4	-	-
27	Hopelessness	-	-	-	0-4	-	-	-	-	-	0-4	-	-
28	Loss of appetite	-	-	-	0-2	-	-	0-2	-	-	-	-	-
29	Weight gain	-	-	-	0-2	-	-	-	0-2	0-2	-	-	-
30	Loss of interest	-	-	-	0-2	-	-	-	-	-	-	-	-
31	Insomnia, general	-	-	-	-	0-4	-	-	-	-	-	-	-
32	Retardation (motor)	-	-	-	-	0-4	-	-	-	-	-	-	-
33	Retardation (verbal)	-	-	-	-	0-4	-	-	-	-	-	-	-

34	Retardation (intellectual)	-	-	-	-	0-4	-	-	-	-	-	-	-
35	Retardation (emotional)	-	-	-	-	0-4	-	-	-	-	-	-	-
36	Loss or gain of weight	-	-	-	-	-	-	0-2	-	-	-	-	-
37	Fatigue	-	-	-	-	-	-	-	0-4	0-4	-	-	-
38	Social withdrawal	-	-	-	-	-	-	-	0-4	0-1	-	-	-
39	Appetite increase	-	-	-	-	-	-	-	0-2	0-3	-	-	-
40	Carbohydrate craving	-	-	-	-	-	-	-	0-3	0-3	-	-	-
41	Hypersomnia	-	-	-	-	-	-	-	0-4	0-4	-	-	-
42	Increased eating	-	-	-	-	-	-	-	-	0-3	-	-	-
43	Diurnal variation Type A	-	-	-	-	-	-	-	-	0-2	-	-	-
44	Diurnal variation Type B	-	-	-	-	-	-	-	-	0-2	-	-	-