Introduction

Observer rating scales (also called observer rating scales, observer scales, or clinical scales) relate to past or current behavior and experiences [1]. These standardized scales are used to rate the extent of psychopathological phenomena and may focus on a single aspect of psychopathology (unidimensional scales) or on several aspects (multidimensional scales). For example, to evaluate only the severity of depression, a unidimensional scale is sufficient. To assess other symptom dimensions, such as anxiety and obsessive-compulsive and psychotic symptoms, a multidimensional scale or a combination of different unidimensional scales is indicated. In a long-term study on patients with depression, a broad range multidimensional scale like the Association for Methodology and Documentation in Psychiatry (AMDp) system [2] might be indicated in order to record mood switches (eg, to manic or psychotic symptoms), despite the focus of the study being on depression [3]. It should be noted that even if the name of a scale appears to indicate that it is unidimensional and focuses on one syndrome or disorder, the scale is very often not actually unidimensional, but multidimensional; this is true for the Hamilton Depression Rating Scale (HAM-D) [4,5].

General aspects of observer rating scales

For each aspect of psychopathology, the assessment may be based on a global rating or on different elements within the aspect being assessed (eg, individual symptoms of the depressive syndrome). In the latter case, the overall score of the instrument is obtained by summing values for these different elements.

Standardization

Assessment or rating scales do not have a uniform level of standardization. Standardization for most of these instruments is limited to providing guidelines that describe the items and the
categories used to assess them and specifying a method to analyze the assessments. Generally, a total score or summary scores (consisting of a total score and subscores) are calculated. For some scales, a time frame and the framework in which the observation should take place are stipulated for the assessment. In the latter case, the instrument is referred to as a fully structured or standardized interview. The more extensive the standardization procedures, the more reliable an assessment instrument generally becomes. However, a highly standardized instrument tends to become less practicable. As a result, for pragmatic reasons the non-fully structured instruments (ie, the typical clinical rating scales such as the HAM-D) are preferred to fully structured instruments in both everyday clinical use and research because they can be completed after a routine psychiatric interview. The inter-rater reliability of these simpler clinical rating scales is lower than that of fully structured assessment instruments. However, this disadvantage can be at least partially compensated for by systematic joint training of raters, as is the case in clinical trials for drug evaluation.

**Symptom assessment**

In observer rating instruments, psychopathological phenomena (symptoms) are identified by trained raters (eg, doctors, psychologists, care staff, lay people trained to administer the instrument) or by relevant others (eg, partner, relatives, friends). The assessment refers to the behavior and/or experience of the patient and is based on the rater’s own observations, information given by the patient or both. Observer rating scales need to be constructed in a way that makes them suitable for the interviewers who will administer them. Thus, some scales are designed for doctors or psychologists trained in psychiatry (eg, Montgomery-Åsberg Depression Rating Scale [MADRS] [6]), while others are designed for care staff trained in psychiatry or for patients’ relatives.

Observer rating scales mainly focus on the psychopathological state. The aim of the scale may be to classify each individual wholly as a ‘case’ or ‘non-case’ [7], record specific aspects of the patient’s mental state [8,9], or assess the whole spectrum of psychopathology (eg, AMDP system). The more syndromes that are represented in a scale, the wider the range of its potential applications will be. For example, a comprehensive multidimensional scale like the AMDP system [2] covers a wide range of symptoms and syndromes characteristic for different mental disorders, including depression. Such a comprehensive scale is useful as part of a clinical basic documentation system that covers all kinds of patients, for example, or in a long-term follow-up study in which patients can be expected to switch into different syndromes (eg, from mania into bipolar depression and vice versa; or from non-psychotic to psychotic major depression and vice versa). However, in order to address specific issues for which even a comprehensive scale does not collect enough data (eg, detailed aspects of suicidal behavior), a comprehensive rating scale should be combined with other specific observer rating scales (eg, the Columbia-Suicide Severity Rating Scale) [10].

When professionally trained assessors administer observer rating instruments, they decide how much weight to put on the information the patient gives. In addition, observable
Changes are taken into account in the rating, for example an improvement in general behavior and demeanor, even if the patient gives no clear report of this improvement. An advantage of this expert assessment is that it reduces the scope for inaccurate assessments resulting from distortions in patients’ perception of themselves. However, it does introduce the risk of distortion related to the assessor (rater bias). Systematic distortion in the assessor’s observations [11] can result from the following factors in particular:

- **Rosenthal effect**: The assessor’s expectations influence the result of the assessment; tendency on the part of the assessor to systematically over- or under-rate the degree of disturbance;
- **Halo effect**: The results of the assessment of one characteristic are influenced by the assessor’s knowledge of the patient’s other characteristics or by the overall impression made by the patient; and/or
- **Logical errors**: The result of the assessment is influenced by assessors reporting only those detailed observations that make sense to them in the context of their theoretical and logical preconceptions. These errors may be partially compensated for by combining observer rating scales with self-rated scales [1].

Most rating scales allow the current mental state to be described. When performed at intervals, they can also be used to examine changes over time, although they were not originally specifically developed for such use. During further development of the scales, changes over time were rather studied with sophisticated statistical analyses focusing on the item development and internal structure of the item association over time and whether the total score of all items or a subset of items always reflect severity in the same way [12,13].

Of interest, especially in the context of psychopharmacological studies, is the administration of observer rating scales like the HAM-D with a telephone-based interactive voice recording system (IVRS), rather than in a face-to-face interview [14–16]. The IVRS can increase reliability and is cost-effective. However, when administered via such a system, the HAM-D is technically no longer being used as a true observer rating scale (ie, although information given by the patient is being assessed, the expert interpretation of this information and clinical observation of depression-related behavior changes such as facial expression are lacking). When used with this new approach, the process becomes similar to a self-rating procedure, with all of its limitations (see Chapter 3).

### Examples of observer rating scales for depression

Several observer rating scales for depression are available; the most traditional and probably the first to be developed, the HAM-D, has been in use for more than 50 years [4,17]. Nearly all antidepressants have been evaluated on the basis of the HAM-D. Although often criticized for several reasons [18], this scale has remained popular for both the evaluation of antidepressants in clinical studies and for other clinical purposes [19]. This scale has an obvious face validity for all doctors trained in psychiatry when they consider its rich coverage of clinically relevant
depression symptoms. In comparison, the MADRS is a more modern scale that has certain advantages such as its shortness (10 items), sensitivity to change, and lack of bias for sedating antidepressants. This lack of bias has resulted in the MADRS being used in many drug trials evaluating modern antidepressants.

Several other depression scales are available [1] that are used under certain conditions or in certain countries, but none are used as widely as the HAM-D and MADRS. Some depression scales were developed in the USA that tended to be primarily based on the symptoms in the Diagnostic and Statistical Manual of Mental Disorders (DSM)-IV description of major depression, for example [20,21]. Although these scales have demonstrated their clinical usefulness in a few larger North American studies such as the Sequenced Treatment Alternatives to Relieve Depression or Systematic Treatment Enhancement Program for Bipolar Disorder, they have not yet become globally accepted. However, the Inventory of Depressive Symptomatology from Rush et al [20] and its respective short version, the Quick Inventory of Depressive Symptomatology [22], might gain wider acceptance in the future. The HAM-D and MADRS are described in detail below. The description includes information on how to use each scale and how it was constructed and psychometrically evaluated, which demonstrates the quality standards of the scale.

In addition to the mental state, domains such as social adjustment may also be measured by observer rating scales. The assessment of social functioning is a useful additional outcome dimension because it helps obtain a full picture of a patient’s problems and burdens. Examples include:

- Social Adjustment Scale (SAS) [23];
- Social Interview Schedule (SIS) [24,25];
- World Health Organization Disability Assessment Schedule (WHODAS) [26];
- Global Assessment Scale (GAS) [27];
- Social and Occupational Functioning Assessment Scale (SOFAS) [28]; and
- Personal and Social Performance Scale [29]

The dimension of social functioning is complementary to the assessment of depressive symptoms; both ratings are only partially intercorrelated, depending on the respective dimension of psychopathology and social functioning [30].

### Featured scale: Hamilton Rating Scale for Depression

The HAM-D [4,5,9] was one of the first observer rating scales for depression to gain worldwide acceptance, although its weaknesses have been increasingly criticized [31,32]. The item scoring sheets of the original 17-item version and the 24-item version of this scale are shown in Tables 2.1 and 2.2 (see Appendix A for the full scale) [33,34].

### How to use the Hamilton Rating Scale for Depression

This observer rating scale is designed to be used by doctors or psychologists trained in psychiatry and with sufficient clinical experience. The rater evaluates the severity of the
symptoms on the basis of information obtained during a clinical interview. Additional information obtained from relatives, friends, nurses and others may also be taken into consideration to enrich or correct the information given by the patient [34]. The interview, which is performed like a typical free or non-standardized psychiatric exploration, should last about 30 minutes to allow time to cover all the relevant points. The scale is intended to measure the severity of symptoms, not minor fluctuations, and therefore the patient’s condition during the past few days or the past week should be considered.

To increase inter-rater reliability and to ensure that the scale is administered correctly (ie, that the items are correctly understood and rated), new users should receive brief training from raters experienced in using the scale. In research studies, either all patients should be assessed by two raters, who should discuss discrepant assessments after the rating, or the whole group of investigators should be given a formal rater training.

The scale measures individual depressive symptoms and their overall severity (reflected in the total score). Sequential HAM-D ratings are often used to assess the course of depression, for example in antidepressant studies. Experience has shown that ratings should generally not

<table>
<thead>
<tr>
<th>Number</th>
<th>Symptom</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>Depressed mood</td>
<td>0–4</td>
</tr>
<tr>
<td>2*</td>
<td>Low self-esteem, guilt</td>
<td>0–4</td>
</tr>
<tr>
<td>3</td>
<td>Suicidal thoughts</td>
<td>0–4</td>
</tr>
<tr>
<td>4</td>
<td>Insomnia: initial</td>
<td>0–2</td>
</tr>
<tr>
<td>5</td>
<td>Insomnia: middle</td>
<td>0–2</td>
</tr>
<tr>
<td>6</td>
<td>Insomnia: late</td>
<td>0–2</td>
</tr>
<tr>
<td>7*</td>
<td>Work and interests</td>
<td>0–4</td>
</tr>
<tr>
<td>8*</td>
<td>Psychomotor retardation</td>
<td>0–4</td>
</tr>
<tr>
<td>9</td>
<td>Psychomotor agitation</td>
<td>0–4</td>
</tr>
<tr>
<td>10*</td>
<td>Anxiety, psychic</td>
<td>0–4</td>
</tr>
<tr>
<td>11</td>
<td>Anxiety, somatic</td>
<td>0–4</td>
</tr>
<tr>
<td>12</td>
<td>Gastrointestinal symptoms (appetite)</td>
<td>0–2</td>
</tr>
<tr>
<td>13*</td>
<td>Somatic symptoms, general</td>
<td>0–2</td>
</tr>
<tr>
<td>14</td>
<td>Sexual disturbances</td>
<td>0–2</td>
</tr>
<tr>
<td>15</td>
<td>Hypochondriasis (somatization)</td>
<td>0–4</td>
</tr>
<tr>
<td>16</td>
<td>Insight</td>
<td>0–3</td>
</tr>
<tr>
<td>17</td>
<td>Weight loss</td>
<td>0–2</td>
</tr>
</tbody>
</table>

| Total score: | 0–53 |

Table 2.1 Scoring sheet for the original, 17-item version of the Hamilton Depression (HAM-D) Rating Scale. The time frame (window) is the past 3 days. *Depression factor. Adapted with permission from Hamilton [4,34] ©BMJ.
Scoring sheet for the 24-item version of the Hamilton Depression (HAM-D) Rating Scale

<table>
<thead>
<tr>
<th>Number</th>
<th>Symptom</th>
<th>Range</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Depressed mood</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Low self-esteem, guilt</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Suicidal thoughts</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Insomnia: initial</td>
<td>0–2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Insomnia: middle</td>
<td>0–2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Insomnia: late</td>
<td>0–2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Work and interests</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Psychomotor retardation</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Psychomotor agitation</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Anxiety, psychic</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Anxiety, somatic</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Gastrointestinal symptoms (appetite)</td>
<td>0–2</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Somatic symptoms, general</td>
<td>0–2</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Sexual disturbances</td>
<td>0–2</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Hypochondriasis (somatization)</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Insight</td>
<td>0–2</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Weight loss</td>
<td>0–2</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Diurnal variation</td>
<td>0–2</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Depersonalization and derealization</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Paranoid symptoms</td>
<td>0–3</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Obsessional and compulsive symptoms</td>
<td>0–2</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Helplessness</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Hopelessness</td>
<td>0–4</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Worthlessness</td>
<td>0–4</td>
<td></td>
</tr>
</tbody>
</table>

**Total score:** (0–75)

Table 2.2 Scoring sheet for the 24-item version of the Hamilton Depression (HAM-D) Rating Scale.
Adapted with permission from Hamilton [4,33] ©BMJ.

be repeated at intervals shorter than 7 days. At repeat interviews, questions about changes in symptoms should be avoided. Also, before interviewing a patient, interviewers should not review the results of the previous rating.

The original scale contains 17 items [4], but 21- and 24-item scales were later developed [33]. Nevertheless, the original 17-item scale is used in and recommended for most psychopharmacological studies. Publications should always give details of the particular version of the scale that was used in a study because the score ranges for some items can vary. If no specific version is named, the reader can assume that the standard version was applied.
The degree of symptom severity is operationally defined for most of the items, which means that the rater must make the assessment on the basis of the content of specific statements and the tone, facial expression and gestures of the patient during the interview; the remaining items depend on a subjective selection of one of a number of levels of severity ranging from ‘absent’ to ‘severe’ or ‘incapacitating’. Most of the items have a three-level score (0–2), the remaining ones a four- (0–3) or five-level (0–4) score, depending on the version of the scale being used. Symptom severity and frequency of symptoms should both be considered in the scoring. The total score of the HAM-D-17 ranges from 0–52 or 53 (depending on the version used), that of the HAM-D-21 from 0–64 and that of the HAM-D-24 from 0–76.

**Quality and characteristics of the Hamilton Rating Scale for Depression**

From a clinical standpoint, the type and spectrum of items characterizing depression seems prima facie meaningful: the items cover the traditional concept of depression, primarily the concept of endogenous depression, and are not influenced by the modern diagnostic classification systems. Somatic symptoms are more broadly represented than affective and cognitive symptoms. Some symptoms such as retardation (item 8) are relatively broad and include cognition, language, and motor activity. Symptoms typical for atypical depression such as hypersomnia and increased appetite or weight cannot be assessed because the scale only asks about changes in the opposite direction (ie, insomnia, reduced appetite, and weight loss). It is questionable whether the characteristic ‘diurnal variations’, which is included in the 21- and 24-item versions, should actually result in a higher depression score. The inclusion of this item can lead to contradictions in the diagnosis of the course of the disease because clinical experience shows that the most severe endogenous depressions often show no diurnal variations at first and that these only occur upon improvement of the severe depressive mood. The fact that three items rate sleep disorders (insomnia early, middle, and late) leads to an efficacy bias in antidepressant studies in favor of sedating or sleep-inducing antidepressants. Experts have discussed whether the slight efficacy advantage of tricyclic antidepressants over selective serotonin reuptake inhibitors (SSRIs) found in some studies and in meta-analyses might have been due to a bias of the HAM-D to capture sleep-inducing, sedating, and anxiolytic properties [35–37].

In addition to the possibility to calculate a total score, factor scores can also be calculated during the final analysis [4]. However, the results of factor analytical evaluations of the scale resulted in different solutions of 2 to 6 factors (which is not unusual for other rating scales in psychiatry) [4,9,38].

The inter-rater reliability is high to very high, at least for the total score, depending on the experience and training of the raters [4,39,40]. The retest reliability is also high [41].

Suggestions were made to define items and assessment criteria more explicitly, in order to increase the inter-rater reliability [42]. Consequently, a Structured Interview Guide was developed for the HAM-D (SIG-H) [41]; the guide has become quite well accepted in Anglo-American countries. However, when the interview guide is used, the HAM-D can no longer be seen as a rating scale but rather as a fully structured interview. As a result, it becomes more
time consuming and no longer fits into the typical communication situation with the patient in a clinical setting and thus induces a somewhat artificial situation.

The HAM-D total score correlates well with the Clinical Global Impression rating of depression and with the total score of other depression scales, indicating its convergent validity [43–45]. As for discriminant validity, the HAM-D correlates moderately well with anxiety scales but the discrimination could be better. However, this finding is similar to that for other depression scales and has been a point of general criticism [46].

The sensitivity of the HAM-D to detect antidepressant-induced changes has been demonstrated in numerous antidepressant studies [47,48] (Figure 2.1) and, recently, in psychotherapy studies [49,50]. The good sensitivity to change can be interpreted as a strong indicator of validity. Reference values for various clinical samples are available [34], but norm values from a representative healthy population are not, as is also the case for most other clinical observer scales. A literature review of control groups in clinical studies of depression reported a mean HAM-D-17 score of 3.2 (SD 3.2) among healthy control individuals [51].

---

**Figure 2.1 Improvement of the Hamilton Depression (HAM-D) Rating Scale total score in severely depressed patients treated with agomelatine**

![Graph showing improvement of HAM-D scores](image)

- Total population: $\Delta = 2.86 \pm 0.56$
- HAM-D $\geq 25$: $\Delta = 3.00 \pm 0.64$
- HAM-D $\geq 25$ and CGI-S $\geq 5$: $\Delta = 3.48 \pm 0.72$
- HAM-D $\geq 30$: $\Delta = 4.53 \pm 1.31$

- **Agomelatine 25–50 mg**
- **Placebo**
- ***$P < 0.001$***

---

**Figure 2.1 Improvement of the Hamilton Depression (HAM-D) Rating Scale total score in severely depressed patients treated with agomelatine.** Dose of 25–50 mg/day for 6–8 weeks (meta-analysis of three positive studies) according to three severity criteria (HAM-D $\geq 25$; HAM-D $\geq 25$ and CGI-S $\geq 5$; HAM-D $\geq 30$). Reproduced with permission from Montgomery and Kasper [48] ©Lippincott, Williams and Wilkins.
Some additional problems of the scale still remain unresolved; for example, it does not record certain diagnostically specific areas that are partially included in other depression scales, such as the Inventory of Depressive Symptomatology (IDS) [28] (which includes symptoms of atypical depression) and, therefore, proves to be unsatisfactory for some subtypes of depression. The IDS is an extended version of the HAM-D and includes all DSM symptoms of major depression. It has 28 items (a more recent version has 30 items) and is available as an observer rating (IDS-C) and self-rated (IDS-SR) scale [52,53].

The HAM-D was subjected to critical test-theoretical analyses, including some performed according to the Rasch model and the facet analytical model, in order to investigate its homogeneity and the stability of the factor structure in repeated measurements during treatment and to find a minimal number of items that adequately reflect severity at different times [20,21,54]. The analyses resulted in a list of unidimensional core items. On the basis of this search for core items with optimal psychometric properties, short versions of the HAM-D were developed to assess depressive symptom severity:

- the Bech six-item version [55];
- a similar six-item version suggested by Maier and Philipp [56]; and
- a seven-item version of the HAM-D (HAM-D-7; [57,58]).

The Bech six-item version showed that this approach might lead to better results than the full-length version of the HAM-D in terms of differentiating antidepressant effects from those of placebo [59]. Other studies also found positive results for the Bech six-item version and other similar short versions [60,61]. Faries et al [62] and Entsuah et al [63] suggested that the use of such an unidimensional short scale requires fewer patients than the full HAM-D-17 scale. Also, such a scale may better detect the real antidepressant effect, independent of sedative and anxiolytic properties of the antidepressant [35]. On the basis of further analyses and clinical reflections, the Bech-Rafaelsen Melancholia Scale was developed (BRMES) [21,64], which consists of 11 items, 6 of which are in the Bech six-item version of the HAM-D.

For pragmatic and other reasons, self-rating versions of the original HAM-D were developed (eg, the Caroll Self-rating Scale for Depression [CDRS] and the Hamilton Depression Inventory [HDI] [65,66]). The latter scale has additional items and increased application possibilities, including a PC version [67] and an interactive voice response (IVR) version [68]. Self-rating versions were also developed from abbreviated versions [33,69].

In addition to reporting mean score changes, drug treatment studies in depression have increasingly focused on using remission as a relevant categorical efficacy criterion. In accordance with Frank et al [70], Rush et al [53] found on the basis of receiver operating characteristic (ROC) analyses of data from patients with major depression, healthy controls, and remitted patients that a HAM-D-17 score ≤7 (which corresponds to a HAM-D-7 score ≤3; [57]) is a meaningful criterion for remission [53]. The ACNP Task Force [71] recommended that if the HAM-D-17 scale is used, a score of ≤7 or ≤5 should be used as the criterion for remission. However, recent evidence supports the use of even more stringent remission criterion scores for both the HAM-D and MADRS scales [72]. A criticism of the use of response (in the
common definition: 50% reduction from the baseline score) as an outcome measure is that it can identify a highly heterogeneous population of patients. However, defining remission with the suggested HAM-D-17 cut-off scores or even more stringent ones identifies populations of remitters that are as heterogeneous as the population of responders in terms of psychosocial impairment [73,74]. Patients with a HAM-D-17 score ≤2, for example, show better psychological functioning than those with scores of 3 to 7 [75,76].

**Featured scale: Montgomery-Åsberg Depression Rating Scale**

Although the HAM-D is still widely accepted and supported by a long tradition and enormous database in terms of psychometric evaluation and repeated use in all kinds of studies, the MADRS [6] is becoming an increasingly important observer rating scale thanks to its conciseness, better definition of items, ease of use, and modern approach to test construction (according to the principle of sensitivity to change). The sensitivity to change aspect seems to support its use in treatment-related studies. Contrary to the HAM-D, which covers a broad spectrum of depressive symptoms, the MADRS includes only the following ten items:

- apparent sadness;
- reported sadness;
- inner tension;
- reduced sleep;
- reduced appetite;
- concentration difficulties;
- lassitude;
- inability to feel;
- pessimistic thoughts; and
- suicidal thoughts.

**How to use the Montgomery-Åsberg Depression Rating Scale**

The scale (Appendix B) should be used by doctors or psychologists trained in psychiatry who have sufficient clinical experience. The ten items are assessed on the basis of a clinical interview and observation. The interview should begin with more general questions and lead on to detailed symptoms. If the patient does not give exact answers, all relevant information from other sources should be integrated into the final evaluation. About 15 minutes seem to be sufficient for the interview. The rating time and length of the interview should be fixed if the scale is used to study the longitudinal course of depressive symptoms.

Each item is rated on a seven-point scale (0–6). Descriptions are given for points 0, 2, 4, and 6 on the scale as anchor points. In the analysis, the scores for each item are summed to give a total score, which can range from 0–60 points. A less expensive version of the scale, which does not include the anchor points, is available and has demonstrated validity and clinical utility [77].
The scale is especially indicated for studies on the clinical course of depression during treatment because it focuses on depression severity and does not cover a broad spectrum of depression symptoms. It is a very economical approach for such settings, given the limited number of items and short duration of the interview. The precise wording of the items guarantees good inter-rater reliability without intensive rater training. Other scales, such as the HAM-D, are preferable if a broader spectrum of symptoms needs to be covered.

**Quality and characteristics of the Montgomery-Åsberg Depression Rating Scale**

The original selection of items was based on the Comprehensive Psychopathological Rating Scale (CPRS) [78], which indicates the content validity. On the basis of frequency analyses and by selecting the items that showed the highest change score and correlated most strongly with the change in the total score, the original number of items related to depression was reduced to ten. The scale includes the main symptoms of depressive illness and most of the DSM criteria for depression, even though certain important areas (eg, psychomotor retardation, tendency to somatize) have been omitted as a result of the method of item selection [79].

As to other aspects of content validity, the scale correlates well with the HAM-D [6,45], especially with the first factor of the HAM-D. Meier et al [46] found a correlation between the MADRS and HAM-D-17 total scores of r=0.85 and between the MADRS and HAM-D-21 total scores of 0.83. Overall, factor analyses and correlations with the HAM-D (particularly with the various subscales) show that the MADRS covers more purely psychological symptoms than the HAM-D [6,79–86]. In these analyses, the dimensions covered by the MADRS items were classified under the headings sadness/pessimistic thoughts, inner tension, inability to feel and reduced appetite. Although the scale does not seem to be unidimensional, it should be noted that more of the MADRS items are loaded on the first factor than are the HAM-D items.

The Rasch model indicated that the MADRS seems to ensure invariance of meaning across different subgroups and also longitudinally [55,80]. As to its discriminant validity, a point of criticism is its only moderate correlation (0.42) with an anxiety scale, the Covi Anxiety Scale [46], and with different subscales of the Positive and Negative Syndrome Scale (PANSS; eg, 0.51 with the PANSS negative subscale) [87].

In the studies performed while the scale was being constructed, the sensitivity to change was claimed to be better than that of other procedures used simultaneously [88–90]. In later studies, the sensitivity of the MADRS for differences in the severity of depression [79] and change in depression symptoms was again shown to be good [80,91,92]. However, when the mean score values of the MADRS were compared with those of the HAM-D in a large sample of inpatients with major depressive disorder, the course appeared to be relatively similar, apart from the higher mean values of the MADRS [93] (Figure 2.2). The effect sizes in placebo-controlled antidepressant efficacy studies were similar to those of the HAM-D-17 scale [94] or slightly better [85,95].

Reference values are available for several clinical samples [80,81], but norm values for the general population are not, as is the case for most scales.
Mean value courses (last observation carried forward [LOCF]) of Hamilton Depression Rating Scale-17 and Montgomery-Åsberg Depression Rating Scale over 10 weeks in 1014 naturalistically treated inpatients with a major depressive episode

![Graph showing mean value courses](image)

Figure 2.2 Mean value courses (last observation carried forward [LOCF]) of Hamilton Depression Rating Scale-17 and Montgomery-Åsberg Depression Rating Scale over 10 weeks in 1014 naturalistically treated inpatients with a major depressive episode. Reproduced with permission from Seemuller et al [93] ©Elsevier.

The inter-rater reliability has been presented for different samples and been shown to be high, with values of 0.89 to 0.97 [6,88]. It was found to be slightly lower in studies that included different professions (psychiatrists, psychologists, nurses) working in psychiatry [80], whereby ratings by nurses showed the lowest inter-rater reliability. The correlation with the Clinical Global Impression of Severity was approximately 0.70 in two studies [45,84]. As mentioned above, drug treatment studies in depression are increasingly using remission as a relevant categorical efficacy criterion. Various suggestions have been made for the remission cut-off score for the MADRS scale: ≤8 [96], <10 [97], or ≤10 [98].

References


Guide to Assessment Scales in Major Depressive Disorder
Alexopoulos, G.; Kasper, S.; Möller, H.-J.; Moreno, C.
2014, IX, 75 p. 3 illus. in color., Softcover
ISBN: 978-3-319-04626-6