Dose-escalation in the picture: pharmacological and imaging studies in depression
Ruhé, H.G.

Citation for published version (APA):
CLINICAL USE OF THE HAMILTON DEPRESSION RATING SCALE: IS INCREASED EFFICIENCY POSSIBLE?
A POST HOC COMPARISON OF HDRS, MAIER AND BECH SUBSCALES, CGI AND SCL-90 SCORES.

Comprehensive Psychiatry 2005; 46: 417-427

Henricus G. Ruhé1,2
Jack J. Dekker1,3
Jaap Peen1
Rebecca Holman4
Frans de Jonghe1,2

1 Mentrum, Amsterdam, The Netherlands
2 Department of Psychiatry, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands
3 Department of Clinical Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
4 Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands
Abstract

Background
The 17-item Hamilton Depression Scale (HDRS) is used as a semi gold standard in research. In treatment guidelines, the HDRS-measurements serve to determine response, remission and guide clinical decision-making for non-responders. However, its use in clinical practice is limited, possibly because the HDRS is time-consuming. Additionally, the multidimensional HDRS is criticized for not measuring a unidimensional aspect as depression severity. The Maier and the Bech, two 6-item severity subscales extracted from the HDRS, are relatively unknown.

Aim
To investigate whether the measurements obtained with Maier and Bech subscales are comparable with the original HDRS-measurements.

Methods
Data from two randomized controlled trials in 482 male and female patients, diagnosed with a major depression (with or without dysthymia) according to DSM-III-R, of whom 219 participated in the trials, were reanalyzed. A standardized stepwise psychopharmacological treatment was compared with a combination of pharmacotherapy with Short Psychodynamic Supportive Psychotherapy in a psychiatric outpatient department. Outcome measures were internal consistency and concurrent validity of HDRS, Maier, Bech, Clinical Global Impression (CGI) scales and Symptom CheckList depression subscale. Effect sizes of HDRS, Maier and Bech were used to compare measured treatment effects for the randomized subjects participating in the trials. Item Response Theory was used to obtain conversion tables for the HDRS, Maier, Bech and SCL-90 depression subscale.

Results
We found moderate internal consistence (Cronbach $\alpha \approx 0.6-0.7$) and high correlations of the Maier and Bech subscales with overall HDRS-scores. Overall there were no clinically relevant differences in effect sizes between Maier, Bech and HDRS, although some differences were statistically significant. Receiver operating characteristic curves showed no difference between Maier and Bech to define remission, but showed the CGI ratings to be unreliable. A cut-off $\leq 4$ corresponded with a HDRS $\leq 7$ criterion in both subscales.

Conclusions
In clinical practice, both Maier and Bech scales can be used as equivalents of the HDRS, but will be more efficient.
Introduction

Major depressive disorder is a severe, disabling illness, expected to be the world’s second health-problem in 2020. Depression is associated with high costs, regarding direct treatment and indirect costs of loss of productivity and quality of life. Several clinical guidelines were developed to guide the treatment of this disorder, both psychotherapy and pharmacotherapy (or in combination) appear effective.

The use of self-report or clinician-rated symptom-scales is recommended to assess severity and response to treatment. Some experts claim clinician-rated symptom scales to have a larger validity and reliability than self-reporting scales, especially in patients with cognitive impairment, and more severe or psychotic depressions. Specific symptom-scales are more reliable than global rating scales. Especially, rating scales can be used to objectively determine specific cut-off points for response and remission.

In most clinical trials the Hamilton Depression Rating Scale (HDRS) – a clinician rated symptom-scale – is used as a standard to determine severity and response. Many versions of the HDRS exist, with the number of items usually varying between 17 and 24, however up to 36 items have been described. Longer versions were especially developed to cover reverse neurovegetative (atypical) symptoms. The Clinical Global Impression (CGI) – a clinician-rated global scale – is also frequently used. In clinical practice, although recommended, rating-scales are not used routinely. Explanations for this discrepancy could be ignorance of existing scales, a strong belief in one’s clinical judgment, an unsystematic approach of depression, but also the amount of time needed for rating-scales (e.g. 15-20 minutes for the HDRS) and the necessity of training.

The HDRS is criticized as being sensitive to somatic symptoms (e.g. somatic illness or side-effects of drugs), for not rating all 9 DSM-IV domains, its unequal weightings of different symptoms and for the multidimensionality of the HDRS total score. Multidimensionality is important to cover the maximum range of clinical features of major depressive disorder, but does not necessarily measures depression severity. Multidimensional scales can be misleading when measurement of severity and treatment response is concerned, especially when the measured depressive symptoms do not change proportionally with depression severity. Finally, some reports emphasize that the HDRS systematically favors (sedative) Tricyclic Antidepressants (TCAs) above Selective Serotonin Reuptake Inhibitors (SSRIs). Sleep and somatic items may appear to be ‘improved’ by side-effects of TCAs, but worsened by side-effects (e.g. insomnia, gastrointestinal complaints, agitation) of SSRIs.

In order to overcome the problems of the multidimensional HDRS mentioned above, a more unidimensional subscale from the HDRS covering core-symptoms of severity is desired. Also, from a clinical point of view, fewer items will be less time consuming for application by busy clinicians. However, for the purpose of reference, subscale scores must remain anchored to the original HDRS. To identify shorter unidimensional subscales, Maier et al. used Rasch- and Mokken-analyses and Gibbons et al. used Factor-analysis. Bech and colleagues developed another 6-item subscale. This scale initially emerged from an analysis with experienced psychiatrists as a validity criterium, and was validated psychometrically thereafter using Rasch-analyses. This Bech subscale was combined with four items of the Cronholm-Ottosson Depression Scale to form the Bech-Rafaelsen Melancholia Scale. Santor and Coyne examined the score-performances of individual HDRS-items as a function of depression severity with a nonparametric Item Response Theory (IRT) approach, retaining 14 items. These 14 items included all 6 items of the Maier subscale and all 8 items of the Gibbons subscale. However one item from the Bech subscale (somatic symptoms) was not included.

In a meta-analysis of individual patient-data, Faries et al. evaluated the responsiveness of total HDRS and subscale scores in TCA and SSRI pharmacotherapy trials, finding a maximal sensitivity for the Maier subscale. In a similar reanalysis, Entsuah et al. found larger effect sizes for the Bech, Maier and Gibbons subscales compared to the HDRS in trials comparing SSRIs or venlafaxine. O’Sullivan et al. found comparable sensitivity to detect changes for the six-item
Bech subscale compared to the 17-item HDRS. Hooper et al. found equal sensitivity to change during treatment for the 6-item Bech subscale compared to the HDRS 17 item version. Möller and Bech et al. used the Bech subscale to reexamine treatment efficacy of SSRIs and mirtazapine (versus TCAs or placebo). The latter publications did not provide data for the Maier subscale.

In this paper we describe a secondary analysis of our trial data, in order to answer the following questions:

1) Are the Maier, Bech and HDRS comparable in the measurement of depression severity and the sensitivity to measure changes in severity? 2) Is this comparability stable across the full range of response to treatment (e.g. non-response, partial and full response), across different treatments and different baseline severity of depression? and 3) What are clinical cut-off points for the subscales to determine remission compared to conventional definitions.

We hypothesized that the differences between Maier, Bech and HDRS-scales would be small and that there would be no apparent effect modification across neither treatments nor baseline severity. In contrast, we hypothesized that for non-responders and partial responders the effect sizes would be smaller than for responders. This would additionally prove the hypothesis of sensitivity to change.

Methods

Patient selection

In the present analyses we use data from two published randomized controlled trials conducted between 1993 and 1998 which were published or accepted for publication. The first trial aimed at efficacy and effectiveness of pharmacotherapy versus the combination of pharmacotherapy with Short Psychodynamic Supportive Psychotherapy (SPSP) (16 sessions). The second trial investigated efficacy and effectiveness of a combination of pharmacotherapy with 8 versus 16 sessions of SPSP. Pharmacotherapy in both trials consisted of three successive steps in case of intolerance or inefficacy. Both trials started with fluoxetine (20 mg/day), when this was unsuccessful (CGI-I >2, only 'minimally improved' or worse) after 6 weeks amitriptyline (≥150 mg/day, dependent of plasma-levels) was initiated in trial 1 and nortriptyline (≥150 mg/day, dependent of plasma-levels) in trial 2. If again unsuccessful after 6 weeks, moclobemide (300-600 mg/day) was started in trial 1 and mirtazapine (30-45 mg/day) in trial 2.

Inclusion criteria for participation in the trials were age between 18 and 60 years, DSM-III-R defined Major Depression (with or without dysthymia) assessed in a structured clinical interview, a 17-item HDRS baseline score of at least 14 points and written informed consent. Patients were excluded in case of psycho-organic or psychotic or dissociative disorders, drug abuse, or when the patient was considered to be too unreliable to participate in a clinical trial. Other axis I comorbidity was not excluded. Further exclusion criteria were if there was a serious communicative or practical problem (e.g. language barrier or the patient will soon leave the country), if there was a contraindication for one of the antidepressants used, if the patient was adequately treated with antidepressants during the present depressive episode, if the patient used other psychotropic medication, or if the patient was or planned to become pregnant. Additional exclusion criteria were of the usual kind in drug research: “too ill” (e.g. antidepressants must be started immediately) and/or “too suicidal” (e.g. hospitalization is unavoidable) to participate in a clinical trial. The study was approved by the medical ethics committee. After complete description of the study to the subjects, written informed consent was obtained.

Of 3226 newly registered outpatients, 988 patients had a depressive disorder. By initial screening 503 of these 988 patients were excluded by the above exclusion-criteria leaving 485 subjects (including patients that later refused to participate or had a HDRS below 14; further referred to as the diagnostic sample). To enter the trials, a second exclusion check was performed by a psychiatrist (excluding 73 patients), and 142 subjects with a HDRS-17 <14 were excluded,
leaving 270 patients for randomization. After randomization 51 patients refused participation, leaving 219 patients who started the proposed therapy (further referred to as the per protocol sample). In this manuscript we used the diagnostic sample for most cross-sectional analyses, and the randomized patients in the per protocol sample for analyses of sensitivity of response-data. For non-completers, the last observation was carried forward (LOCF).

**Outcome measures**

Primary outcome measures were the 17-item HDRS, the Maier subscale of the HDRS (containing items 1, 2, 7-10), the Bech subscale of the HDRS (items 1, 2, 7, 8, 10 and 13), the Clinical Global Impression Severity (CGI-S) and Improvement (CGI-I) scale, and the Symptoms check list, 90 items (SCL-90) depression subscale (SCL-90 dep). Thus, three levels of information were obtained: data from 1) an independent, trained, supervised and blinded research assistant (HDRS & Maier, Bech), 2) the treating clinician (CGI-S/I) and 3) the patient (SCL-90 dep). The HDRS was administered using a semi-structured interview. Before participating in the study, the reliability of the HDRS-assessments was established. During the study, in order to avoid slippage, audiotaped assessments were discussed monthly.

In the analyses of treatment efficacy, response was defined as a ≥50% HDRS-score reduction, partial response as ≥20-50% reduction in HDRS-score and remission as a HDRS score of 7 points or less. In the analyses of treatment efficacy, response was defined as a ≥50% HDRS-score reduction, partial response as ≥20-50% reduction in HDRS-score and remission as a HDRS score of 7 points or less.

**Statistics**

Cronbach-α coefficients and mean inter-item correlations were used to express internal consistency. To check whether the increased number of items in the HDRS accounted for a higher Cronbach-α coefficient than in the subscales (with only 6 items), we applied the Spearman-Brown formula. Next we calculated concurrent validity as Pearson correlation coefficients between total HDRS, Maier and Bech subscale scores and SCL-90 dep scores. Linear regression models calculated variance of HDRS-scores explained by the subscales. These analyses were performed in our diagnostic sample. Concurrent validity between CGI-S/I and HDRS subscale ratings was determined also, however, to avoid low correlations due to limited dispersion, this was done for the last observation in the per protocol sample. The CGI improvement scale was compared with changes expressed as percentages of the baseline score.

In order to compare differences in sensitivity to measure treatment effects (also referred as responsiveness), in data from the per protocol sample effect sizes (E-S) for HDRS, Maier and Bech subscales were calculated per subject as the within-subject changes in scale-scores divided by the pooled standard deviation of the mean change in scale score. In this way, differences in effect sizes could be tested and 95% confidence intervals (95% CI) could be calculated. Differences in E-S between the scales were tested by paired t-tests. In order to determine significant effect-modification, the above analyses were repeated while data were stratified. For stratification we used initial HDRS scores of at least 19 for severe depression, criteria for response as described above and treatment-condition. Differences in E-S between strata were tested by Analysis of variance-models (ANOVA). The Partial Credit Item Response Theory (IRT) model was used to estimate the relationships between total scores on the HDRS and total scores on the Maier and Bech subscales of the HDRS. The scores were those obtained at exit (per protocol sample). The computer program OPLM was used to obtain a set of weights for each item in the HDRS using conditional maximum likelihood methods. The same software and the item weights were used to obtain estimates of the latent trait associated with each score on the HDRS, the Maier subscale and the Bech subscale. The total scores for the pairs of scales were equated by matching the total scores for which the latent trait scores were most similar. These methods are very similar to those used in a previous publication about the Quick Inventory for Depressive Symptomatology IDS. The range of SCL scores associated with each HDRS score was obtained directly from the original data.
Finally, Receiver operator characteristic (ROC) curves were constructed to summarize validity of cut-off points. Differences in areas under the curve (AUC) were tested with attention for interrelation (because we studied these tests within the same subjects) as described by Hanley.\textsuperscript{61} For all data analysis except the IRT analysis, SPSS for Windows version 10.1 was used.\textsuperscript{62} For all tests two-tailed significance levels were applied.

Table 3.1. Studied populations.*

<table>
<thead>
<tr>
<th></th>
<th>Diagnostic sample (n= 485)\textsuperscript{a}</th>
<th>Per protocol sample (n= 219)\textsuperscript{b}</th>
<th>Combined I + II</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex (%) female</strong></td>
<td>60.3</td>
<td>63.2</td>
<td>61.1</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>35.3 ±9.9</td>
<td>34.9 ±8.2</td>
<td>34 ±9.4</td>
</tr>
<tr>
<td><strong>Marital status (%)</strong>:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>married</td>
<td>97 (20.2)</td>
<td>12 (21.1)</td>
<td>10 (13.9)</td>
</tr>
<tr>
<td>divorced</td>
<td>60 (12.5)</td>
<td>7 (12.3)</td>
<td>8 (11.1)</td>
</tr>
<tr>
<td>widowed</td>
<td>3 (0.6)</td>
<td>-</td>
<td>1 (2.2)</td>
</tr>
<tr>
<td>never married</td>
<td>38 (66.1)</td>
<td>38 (66.7)</td>
<td>54 (75.0)</td>
</tr>
<tr>
<td>other</td>
<td>3 (0.6)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Educational level (%)</strong>:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>72 (15)</td>
<td>11 (19.3)</td>
<td>13 (18.3)</td>
</tr>
<tr>
<td>intermediate</td>
<td>179 (37.3)</td>
<td>21 (36.8)</td>
<td>22 (31.0)</td>
</tr>
<tr>
<td>high</td>
<td>229 (47.7)</td>
<td>25 (43.9)</td>
<td>36 (50.7)</td>
</tr>
<tr>
<td><strong>Occupational (%)</strong>:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>job</td>
<td>166 (34.7)</td>
<td>18 (31.6)</td>
<td>22 (30.6)</td>
</tr>
<tr>
<td>on sickness</td>
<td>134 (28.0)</td>
<td>14 (24.6)</td>
<td>23 (31.9)</td>
</tr>
<tr>
<td>social security</td>
<td>84 (17.5)</td>
<td>12 (21.1)</td>
<td>10 (13.9)</td>
</tr>
<tr>
<td>disabled</td>
<td>27 (5.6)</td>
<td>3 (5.3)</td>
<td>5 (6.9)</td>
</tr>
<tr>
<td>student</td>
<td>41 (8.6)</td>
<td>5 (8.8)</td>
<td>10 (13.9)</td>
</tr>
<tr>
<td>other</td>
<td>27 (5.6)</td>
<td>5 (8.8)</td>
<td>2 (2.8)</td>
</tr>
<tr>
<td><strong>Duration of episode (&lt; 1 year)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-2 years</td>
<td>314 (67.0)</td>
<td>39 (70.9)</td>
<td>49 (70.0)</td>
</tr>
<tr>
<td>&gt; 2 years</td>
<td>70 (14.9)</td>
<td>6 (10.9)</td>
<td>9 (12.9)</td>
</tr>
<tr>
<td><strong>Psychiatric treatment during this episode (%)</strong>:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>antidepressants ≤3 mths before (%)</td>
<td>77 (16.4)</td>
<td>6 (10.9)</td>
<td>10 (14.3)</td>
</tr>
<tr>
<td>adequate (%)</td>
<td>1 (0.2)</td>
<td>-</td>
<td>9 (20.5)</td>
</tr>
<tr>
<td>Depressive episodes (prev. 5 years)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>296 (63.1)</td>
<td>26 (47.3)</td>
<td>4 (58.6)</td>
</tr>
<tr>
<td>1-2</td>
<td>138 (29.4)</td>
<td>21 (38.2)</td>
<td>22 (31.4)</td>
</tr>
<tr>
<td>3 or more</td>
<td>35 (7.5)</td>
<td>8 (14.5)</td>
<td>7 (10.0)</td>
</tr>
<tr>
<td><strong>Ethnicity (%)</strong>:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North-west Europe</td>
<td>414 (86.3)</td>
<td>51 (89.5)</td>
<td>63 (87.5)</td>
</tr>
<tr>
<td>Mediterranean</td>
<td>18 (3.8)</td>
<td>3 (5.3)</td>
<td>2 (2.8)</td>
</tr>
<tr>
<td>Caribbean</td>
<td>22 (4.6)</td>
<td>1 (1.8)</td>
<td>2 (2.8)</td>
</tr>
<tr>
<td>Other</td>
<td>26 (5.4)</td>
<td>2 (3.5)</td>
<td>5 (6.9)</td>
</tr>
<tr>
<td><strong>Baseline scores of rating scales</strong>:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDRS-17</td>
<td>17.1 ±6.5</td>
<td>21.0 ±4.8</td>
<td>20 ±4.9</td>
</tr>
<tr>
<td>Maier</td>
<td>9.2 ±3.6</td>
<td>11.0 ±2.9</td>
<td>10.9 ±2.8</td>
</tr>
<tr>
<td>Bech</td>
<td>9.4 ±3.7</td>
<td>11.5 ±2.8</td>
<td>11.2 ±2.7</td>
</tr>
<tr>
<td>CGI-S \textsuperscript{c}</td>
<td>4.7 ±0.7</td>
<td>4.8 ±0.6</td>
<td>4.7 ±0.7</td>
</tr>
<tr>
<td>SCL-90 depression subscale</td>
<td>45.9 ±11.8</td>
<td>48.7 ±11.7</td>
<td>47.8 ±9.8</td>
</tr>
</tbody>
</table>

* Data represent means (±SD) unless indicated. Denominators of percentages vary due to missing values.
\textsuperscript{a} Total diagnostic sample.
\textsuperscript{b} No significant differences between treatment-groups (per protocol sample) (ANOVA or \(\chi^2\)).
\textsuperscript{c} Significant differences (p < 0.05) between in- and excluded patients (indep. T-Test).
\textsuperscript{d} \(n= 241\)

References to studies: Trial I\textsuperscript{45} and Trial II\textsuperscript{46}
Results

Patient characteristics
Table 3.1 shows demographics for the diagnostic and per protocol samples. There were no significant differences observed between the diagnostic and per protocol sample (tested as excluded versus included), except from a lower mean HDRS-score (and Maier, Bech and SCL90 depression scores) in the diagnostic sample. This difference was due to application of the entrance criterion (HDRS ≥14) for randomization. No significant differences existed between the different treatment-groups. The studied population existed of mainly unmarried, mid-thirty, moderately to highly educated, female, Caucasian adults, with moderate to severe depressive episodes of less than 1 year of duration. More than 75% of the subjects were not treated for the current depressive episode before, 16% percent received an inadequate trial of an antidepressant.

Table 3.2. Internal validity and concurrent validity of HDRS-17, Maier, Bech, SCL-90 depression subscale and CGI-S.

<table>
<thead>
<tr>
<th>Internal consistency</th>
<th>Concurrent validity: Pearson’s r (% explained variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach’s α</td>
<td>Mean inter-item corr.</td>
</tr>
<tr>
<td>Maier†</td>
<td></td>
</tr>
<tr>
<td>Diagnostic sample</td>
<td></td>
</tr>
<tr>
<td>Maier</td>
<td>0.62</td>
</tr>
<tr>
<td>HDRS-17†</td>
<td></td>
</tr>
<tr>
<td>SCL-90 depression subscale†</td>
<td>0.88</td>
</tr>
<tr>
<td>Per protocol sample</td>
<td></td>
</tr>
<tr>
<td>CGI-S endpoint‡</td>
<td>NA</td>
</tr>
<tr>
<td>CGI-I endpoint§</td>
<td>NA</td>
</tr>
<tr>
<td>Per protocol sample</td>
<td></td>
</tr>
<tr>
<td>CGI-S: n= 482; SCL-90dep: n= 473.</td>
<td></td>
</tr>
</tbody>
</table>

* Severe depression defined as initial HDRS-17 ≥19 (n= 221).
† Maier, Bech, HDRS: n= 482; SCL-90dep: n= 473.
‡ CGI-S: n= 229.
§ Compared with change expressed as percentage of baseline rating

Internal and concurrent validity
Data for internal and concurrent validity are presented in Table 3.2. Cronbach-alphas were slightly lower for the Maier and Bech subscales. If a 17-item scale is reduced to 6 items, the expected alpha is 0.49 (Spearman-Brown formula). Thus the observed values of 0.62 and 0.60 show increased internal validity for the subscales. The mean inter-item correlation was markedly higher for the Maier and Bech subscales. The correlation between Maier and Bech subscales was high. Both Maier and Bech subscales explained about 75% of the variance of the total HDRS-score. The self-rated SCL-90dep was reasonably well correlated with the HDRS (r= 0.67) and the Maier and Bech subscales (r= 0.64). Concurrent validity of the scales was overall slightly less in the more depressed sub-group (HDRS ≥19; n= 194) compared to moderately depressed subjects, except for the correlation between HDRS and Maier subscale. The CGI-S at study-endpoint was moderately correlated with the HDRS (r= 0.57), as with the Maier and Bech subscales. The CGI-S showed higher correlation with the Bech subscale, especially in those severely depressed. The CGI-I at study-endpoint was less well correlated with the percentage change in HDRS (r= 0.42) and the subscales.
Sensitivity to change

In Table 3.3 and 3.4, overall and stratified E-S in the per protocol sample are presented. In these tables the 95% CI of the E-S indicates whether the E-S significantly deviates from 0 (no effect measured). Comparisons between E-S may produce significant differences between E-S, even when the 95% CIs between the two E-S overlap.

Of the 9 comparisons between the Maier and Bech subscales made in these tables, 5 were not significant. The Maier was significantly different from the HDRS in 4 out of 9 comparisons, while the Bech was significantly different from the HDRS in only 1 of the 9 comparisons. Differences between E-S were small.

Table 3.3. Pre- and post-treatment Maier, Bech and HDRS-scores with corresponding effect sizes in per protocol sample. Stratification by depression-severity and final treatment response.

<table>
<thead>
<tr>
<th></th>
<th>Mean ±SD baseline</th>
<th>Mean ±SD endpoint (LOCF)</th>
<th>Mean decrease (95% CI)</th>
<th>SD of decrease</th>
<th>Effect size [E-S] (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All subjects (n= 219)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maier</td>
<td>10.9 ±2.75</td>
<td>6.2 ±4.46</td>
<td>4.7 (4.1 - 5.3)</td>
<td>4.54</td>
<td>1.03 (0.89 - 1.16)†</td>
</tr>
<tr>
<td>Bech</td>
<td>11.1 ±2.69</td>
<td>6.2 ±4.50</td>
<td>4.9 (4.3 - 5.5)</td>
<td>4.54</td>
<td>1.08 (0.95 - 1.22)</td>
</tr>
<tr>
<td>HDRS-17</td>
<td>20.2 ±4.56</td>
<td>12.0 ±7.62</td>
<td>8.2 (7.2 - 9.2)</td>
<td>7.45</td>
<td>1.10 (0.96 - 1.23)</td>
</tr>
<tr>
<td><strong>Moderate depression (initial HDRS-17 &lt;19; n= 93)</strong>§</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maier</td>
<td>8.9 ±2.13</td>
<td>5.3 ±3.86</td>
<td>3.7 (2.8 - 4.5)</td>
<td>4.19</td>
<td>0.81 (0.62 - 1.00)*</td>
</tr>
<tr>
<td>Bech</td>
<td>9.2 ±2.12</td>
<td>5.1 ±3.89</td>
<td>4.1 (3.2 - 5.0)</td>
<td>4.28</td>
<td>0.91 (0.71 - 1.10)</td>
</tr>
<tr>
<td>HDRS-17</td>
<td>16.2 ±1.42</td>
<td>9.8 ±6.15</td>
<td>6.4 (5.1 - 7.8)</td>
<td>6.47</td>
<td>0.86 (0.68 - 1.04)</td>
</tr>
<tr>
<td><strong>Severe depression (initial HDRS-17 ≥19; n= 126)</strong>‡</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maier</td>
<td>12.3 ±2.25</td>
<td>6.9 ±4.75</td>
<td>5.4 (4.6 - 6.2)</td>
<td>4.66</td>
<td>1.19 (1.01 - 1.37)*</td>
</tr>
<tr>
<td>Bech</td>
<td>12.5 ±2.16</td>
<td>7.0 ±4.75</td>
<td>5.5 (4.7 - 6.3)</td>
<td>4.65</td>
<td>1.21 (1.03 - 1.39)</td>
</tr>
<tr>
<td>HDRS-17</td>
<td>23.1 ±3.79</td>
<td>13.7 ±8.18</td>
<td>9.5 (8.1 - 10.8)</td>
<td>7.88</td>
<td>1.27 (1.08 - 1.46)</td>
</tr>
<tr>
<td><strong>Final Non-responders (n= 65)</strong>§</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maier</td>
<td>10.6 ±2.82</td>
<td>10.6 ±2.86</td>
<td>0.0 (-0.7 - 0.6)</td>
<td>2.65</td>
<td>-0.01 (-0.15 - 0.14)†</td>
</tr>
<tr>
<td>Bech</td>
<td>11.1 ±2.76</td>
<td>10.7 ±2.89</td>
<td>0.4 (-0.2 - 1.1)</td>
<td>2.66</td>
<td>0.09 (-0.05 - 0.24)§</td>
</tr>
<tr>
<td>HDRS-17</td>
<td>19.5 ±4.25</td>
<td>19.9 ±4.55</td>
<td>-0.4 (-1.2 - 0.3)</td>
<td>3.08</td>
<td>-0.06 (-0.16 - 0.05)¶</td>
</tr>
<tr>
<td><strong>Final Partial-responders (n= 64)</strong>§</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maier</td>
<td>11.5 ±2.86</td>
<td>7.4 ±3.07</td>
<td>4.1 (3.4 - 2.7)</td>
<td>2.58</td>
<td>0.90 (0.76 - 1.04)</td>
</tr>
<tr>
<td>Bech</td>
<td>11.4 ±2.59</td>
<td>7.4 ±3.19</td>
<td>4.1 (3.4 - 2.7)</td>
<td>2.71</td>
<td>0.90 (0.75 - 1.04)</td>
</tr>
<tr>
<td>HDRS-17</td>
<td>21.4 ±5.06</td>
<td>14.2 ±4.41</td>
<td>7.2 (6.7 - 7.7)</td>
<td>1.88</td>
<td>0.96 (0.90 - 1.03)</td>
</tr>
<tr>
<td><strong>Final Responders (n= 90)</strong>§</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maier</td>
<td>10.6 ±2.59</td>
<td>2.1 ±2.00</td>
<td>8.5 (7.8 - 9.1)</td>
<td>3.12</td>
<td>1.87 (1.72 - 2.01)**</td>
</tr>
<tr>
<td>Bech</td>
<td>10.9 ±2.71</td>
<td>2.1 ±1.89</td>
<td>8.8 (8.1 - 9.4)</td>
<td>3.14</td>
<td>1.93 (1.79 - 2.08)</td>
</tr>
<tr>
<td>HDRS-17</td>
<td>19.9 ±4.28</td>
<td>4.8 ±3.46</td>
<td>15.1 (14.1 - 16.1)</td>
<td>4.89</td>
<td>2.02 (1.89 - 2.16)**</td>
</tr>
</tbody>
</table>

Note that the overlap of two 95%CI of E-S does not rule out a statistical significant difference between these E-S (see text).

* Significantly different from E-S_HDRS (paired T-test; p < 0.05)
† Significantly different from E-S_Bech (paired T-test; p < 0.05)
‡ Significant differences of E-S_Maier, E-S_Bech and E-S_HDRS between moderate and severe depression (ANOVA; p< 0.05)
§ Response criteria: decrease in HDRS-scores: <20%= Non-response, 20-50%= Partial response and ≥50%= Response. Significant differences of E-S_Maier, E-S_Bech and E-S_HDRS between categories of response (ANOVA; p< 0.001)
¶ Significant differences between E-S_Maier-E-S_Bech, E-S_HDRS-E-S_Maier and E-S_HDRS-E-S_Bech (paired T-test; p < 0.05)** Significant difference between E-S_HDRS-E-S_Maier (paired T-test; p < 0.05)
Clinical use of HDRS subscales

In the total per protocol sample the Maier subscale was significantly less powerful to observe treatment effects: the E-S assessed by the Maier was significantly lower than the E-S of the Bech and HDRS. When stratified for depression severity, the E-S of Maier, Bech subscales and HDRS were larger in severe compared to moderate depression. A significant difference between these strata was observed for all E-S (ANOVA).

Within the group of severely depressed subjects, the Maier was significantly less sensitive compared to the HDRS (paired T-test). Within the moderately depressed group, the Bech outperformed the Maier (paired T-test). Across different strata of final response significant differences in E-S were found (ANOVA). Within strata, the Bech subscale performed less in final non-responders, while the Maier performed significantly less than the total HDRS in final responders (paired T-tests).

In Table 3.4 it is shown that no overall differences in effect sizes were found between treatment modalities (ANOVA). Within the group of patients treated with anti-depressants only, the Maier subscale was significantly less sensitive to detect treatment differences than the HDRS, however the Maier did not differ significantly from the Bech subscale (paired T-test).

Conversion of HDRS scores, criteria for remission and depression severity

Table 3.5 shows the conversion between HDRS scores and Maier, Bech subscales and SCL90 scores. Maier and Bech cut-off scores to define remission (bold), mild, moderate and severe depression can be identified (italic). Figure 3.1 shows the ROC-curves for Maier, Bech CGI-S, CGI-I and SCL90 cut-off scores, with HDRS ≤7 as the reference criterion. The difference in AUC for the Maier and Bech subscales was not significant (z= 1.25; p= 0.21). The difference in AUC between Maier and Bech subscales compared to SCL90 and both CGIs was highly significant (z> 3.8; p< 0.001). In the table below Table 3.1 sensitivity and specificity for cut-off scores ≤3 and ≤4 for the Maier and Bech subscales are given.
Figure 3.1. Receiver operating characteristic curves for Maier and Bech subscales, SCL90 depression and CGI-S/I at endpoint compared to HDRS-17 'Remission' in per protocol sample.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-scale ≤3</td>
<td>88.9</td>
<td>93.2</td>
<td>87.5</td>
<td>92.5</td>
</tr>
<tr>
<td>Sub-scale ≤4</td>
<td>97.2</td>
<td>85.7</td>
<td>97.2</td>
<td>83.7</td>
</tr>
</tbody>
</table>

AUC (SE): Maier 0.972 (.009), Bech 0.963 (.011), SCL-90dep 0.862 (.028), CGI-S 0.743 (.036), CGI-I 0.738 (.035)

Discussion

Major findings

This study examined the relative effectiveness of the HDRS subscales as developed by Maier et al. and Bech et al. in monitoring severity and treatment effects in depression. We found that the Maier and Bech subscales gave results comparable to the original 17-item HDRS, with high concurrent validity and increased mean inter-item correlations and internal consistency. Maier and Bech subscales were highly comparable to each other in the measurement of treatment changes. Differences between E-S were rather small, and clinically irrelevant. For interpretation a conversion table linking HDRS-scores and Maier and Bech scores are provided. The Maier had a slightly (non-significant) higher sensitivity and specificity to predict the reference criterion for remission (HDRS ≤7). Both Maier and Bech subscales differentiated non-responders from partial and final responders.

A significant difference in sensitivity to change existed between the Bech and Maier within the group treated with antidepressants only. We were unable to find the reason for this difference compared to other treatment modalities, where the difference between Maier and Bech was not found or was not significant. The question arises whether there is a difference in sensitivity between the Maier and Bech subscales across different treatment-modalities or that other (post-randomization) differences between the groups or mere chance explain this observation. Because this difference was not found in the other groups (treated with both antidepressants and psychotherapy) we think it cannot be ascribed to a difference in detecting pharmacological (side-) effects. If a Bonferroni correction would be applied for the number of comparisons tested (p < 0.01), the observed difference would not remain its significance.
The relevance of the difference between the Maier and Bech subscales

The only difference between the Maier and Bech subscales is the inclusion of agitation (e.g. running thoughts or restlessness, 0-4 points) in the Maier versus the inclusion of general somatic symptoms (e.g. tiredness, 0-2 points) in the Bech. It could be argued that one scale is comparable with the other scale without the different item, e.g. the Maier subscale would then be comparable to the Bech subscale minus the ‘general somatic’ item. In our (diagnostic) sample the item agitation contributed 1.3 (SD= 0.9) points to the total Maier-score (9.2, SD= 3.6). The general somatic item contributed 1.5 (SD= 0.8) points to the total Bech-score (9.4, SD= 3.7). Thus, overall

Table 3.5. The conversion between the HDRS total scores and the Maier subscale, Bech subscale and the range of SCL scores using IRT analysis (per protocol sample).*

<table>
<thead>
<tr>
<th>HDRS</th>
<th>Maier</th>
<th>Bech</th>
<th>Range SCL90_{dep}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>1 – 2</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>3 – 4</td>
<td>2</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>7 (remission†)</td>
<td>4</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>8 – 9</td>
<td>5</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>10 – 11</td>
<td>6</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>13 (mild†)</td>
<td>7</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>8</td>
<td>31 – 61</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>9</td>
<td>22–59</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
<td>9</td>
<td>26 – 61</td>
</tr>
<tr>
<td>17</td>
<td>9</td>
<td>10</td>
<td>25–59</td>
</tr>
<tr>
<td>18 (moderate†)</td>
<td>10</td>
<td>10</td>
<td>30–62</td>
</tr>
<tr>
<td>19</td>
<td>10</td>
<td>10</td>
<td>28 – 60</td>
</tr>
<tr>
<td>20</td>
<td>11</td>
<td>11</td>
<td>38 – 67</td>
</tr>
<tr>
<td>21</td>
<td>11</td>
<td>11</td>
<td>30 – 72</td>
</tr>
<tr>
<td>22</td>
<td>12</td>
<td>12</td>
<td>39 – 61</td>
</tr>
<tr>
<td>23</td>
<td>12</td>
<td>12</td>
<td>36 – 61</td>
</tr>
<tr>
<td>24 (severe†)</td>
<td>13</td>
<td>13</td>
<td>38 – 67</td>
</tr>
<tr>
<td>25 (very severe†)</td>
<td>13</td>
<td>13</td>
<td>45 – 64</td>
</tr>
<tr>
<td>26</td>
<td>13</td>
<td>13</td>
<td>47 – 71</td>
</tr>
<tr>
<td>27</td>
<td>14</td>
<td>14</td>
<td>46 – 72</td>
</tr>
<tr>
<td>28</td>
<td>14</td>
<td>14</td>
<td>42 – 79</td>
</tr>
<tr>
<td>29</td>
<td>15</td>
<td>14</td>
<td>55 – 71</td>
</tr>
<tr>
<td>30</td>
<td>15</td>
<td>15</td>
<td>43 – 43</td>
</tr>
<tr>
<td>31</td>
<td>15</td>
<td>15</td>
<td>63 – 70</td>
</tr>
<tr>
<td>32</td>
<td>16</td>
<td>15</td>
<td>62 – 65</td>
</tr>
<tr>
<td>33</td>
<td>16</td>
<td>16</td>
<td>57 – 71</td>
</tr>
<tr>
<td>34 – 35</td>
<td>17</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>36</td>
<td>17</td>
<td>17</td>
<td>-</td>
</tr>
<tr>
<td>37 – 39</td>
<td>18</td>
<td>17</td>
<td>-</td>
</tr>
<tr>
<td>40 – 42</td>
<td>19</td>
<td>18</td>
<td>-</td>
</tr>
<tr>
<td>43 – 44</td>
<td>20</td>
<td>19</td>
<td>-</td>
</tr>
<tr>
<td>45</td>
<td>21</td>
<td>19</td>
<td>-</td>
</tr>
<tr>
<td>46</td>
<td>21</td>
<td>20</td>
<td>-</td>
</tr>
<tr>
<td>47 – 48</td>
<td>22</td>
<td>20</td>
<td>-</td>
</tr>
</tbody>
</table>

* The only valid conversions that can be made from this table are between 1) HDRS and Maier, 2) HDRS and Bech and 3) HDRS and SCL90_{dep}.
† Cut-offs as provided by Yonkers.11

The relevance of the difference between the Maier and Bech subscales

The only difference between the Maier and Bech subscales is the inclusion of agitation (e.g. running thoughts or restlessness, 0-4 points) in the Maier versus the inclusion of general somatic symptoms (e.g. tiredness, 0-2 points) in the Bech. It could be argued that one scale is comparable with the other scale without the different item, e.g. the Maier subscale would then be comparable to the Bech subscale minus the ‘general somatic’ item. In our (diagnostic) sample the item agitation contributed 1.3 (SD= 0.9) points to the total Maier-score (9.2, SD= 3.6). The general somatic item contributed 1.5 (SD= 0.8) points to the total Bech-score (9.4, SD= 3.7). Thus, overall
tiredness was more present than agitation in this sample, and agitation was not rated near its maximum like tiredness. Both items occurred intraindividually at the same time, but were not interchangeable. This means that the Maier and Bech subscales show different perspectives on depressive symptomatology. In this respect, it is noteworthy to mention that the agitation item was dropped beforehand when the Bech subscale was developed and validated, because this item showed limited variance (i.e. was not found to be scored) in the two studied samples. Furthermore, in the Maier subscale the items psychomotor agitation and psychomotor retardation are included, which -at first sight- seem to represent two opposed polarities. However, these items also co-occurred within the same individuals. This can be explained by the broad definitions of agitation (both restlessness or running thoughts) and retardation (both retardation in activities or in thinking) in our semi-structured interview.

The original HDRS is often criticized to measure these somatic symptoms. Although the Bech subscale was designed as an unidimensional scale, the 'general somatic' item is still among the 6 items. However, in the Rasch-analysis this item was the least contributive and showed a ceiling effect for moderate and severe depression. Though both a DSM-IV and ICD-10 criterion for diagnosis, the aspecific 'tiredness' symptom may also be caused by physical illnesses. The Maier subscale does not include this item. Thus, the Maier subscale might especially be useful in patients suffering from somatic complaints or illness. Additional, methodological support comes from the exclusion of this somatic item by Santor et al. This hypothesis of better performance in patients with somatic complaints or illness needs further investigation, e.g. in comparison with the Hospital Anxiety and Depression Scale.

Previous comparative studies

Our findings are in line with findings of previous studies, and extends the evidence to support the Maier and Bech subscale as a valid alternative for the HDRS. This is relevant for the planning and conduction of clinical trials, but also for clinical practice. Hooper et al. found equal performance of the Bech subscale compared to the Montgomery Asberg Depression Rating Scale (MADRS). Because the MADRS was not used in our trials, we were unable to examine the performance of the Maier subscale compared to the MADRS. Hooper et al. questioned whether a possible ceiling effect in the Bech subscale would limit its usefulness in severely depressed patients. In our study more than 57% of the per protocol patients had an initial HDRS greater than 18 (indicative for severe depression). We did not find a ceiling effect in our diagnostic sample (data not shown), and found consequently higher effect sizes for the Maier and Bech subscales in initially severely depressed patients, indicative for an adequate sensitivity to measure (larger) changes due to treatment. In addition to the observed ability to predict remission, we proposed cut-off scores for remission and various ranges for classification of depression severity.

In two publications Bech proposed the Bech subscale as an alternative measure to overcome the confounding influence of drug related side-effects in the comparison with placebo or active drugs. However this problem is not fully solved, as tiredness may be induced by histaminergic effects from antidepressants (e.g. tricyclics and mirtazapine). On the other hand, agitation (included in the Maier subscale) is known as a (mostly transient) SSRI-induced side-effect.

An extra dimension of our study is that it extends the data for use of the Maier and Bech subscales in populations treated with psychotherapy. Hooper et al. and O’Sullivan et al. already demonstrated the usefulness of the Bech subscale in pharmacological treatment of melancholia, dysthymia and typical and atypical depression.

An alarming point of our study is the moderate correlation of the Maier, Bech and HDRS with the CGI-S and the CGI-I. Previous reports mentioned correlations between HDRS and CGI varying between 0.65 and 0.90. Our results underscore the need of a HDRS or subscale rating instead of the CGI. We consider the validity of the CGI to be questionable, as most CGI-raters (subjectively) evaluate their own treatment. Apparently, the clinician’s judgment does not coincide with scale scores. In this respect, the performance of the (self-rated) SCL-90dep is better. This was also illustrated in the ROC-curves regarding the criterion of remission. Further research is needed to investigate whether correlations with the HDRS of other self-rated scales (e.g. the Beck
Depression Inventory\textsuperscript{64}) are higher than the SCL-90\textsubscript{dep}. In addition to this, a major limitation in our study and in any study investigating depression severity, is that there is no definite gold standard. We used HDRS data as the gold standard, which means that scales under investigation can never be judged to be better than the HDRS, however this would be reversed if the CGI was used as a gold standard.\textsuperscript{65}

**Conclusion**

We conclude that both Maier and Bech subscales of the HDRS are equivalent to the HDRS, and can easily be used to increase efficiency to measure treatment response in clinical practice. On theoretical grounds we have a slight preference for the Maier subscale. The use of subscales would improve the efficiency and objectivity of measuring response in clinical practice, where often no scale (instead of a Clinical Global Impression) is used at all. This would further bridge the gap between clinical practice and research based treatment-recommendations for non-response in depression. Maier and Bech subscales should be compared in patients suffering from co-morbid somatic illnesses, or patients treated with psychotherapy only. The impact of the difference of the one somatic item versus the agitation-item between the Maier and Bech subscales and the consequences for their applicability in clinical sub-groups needs further research.

**Acknowledgement**

The original randomized controlled trials were supported by an unrestricted educational grant from Eli Lilly Netherlands. All studies were carried out by the Mentrum Depression Research Group. We thank all psychotherapists, psychiatrists and residents for their excellent work.

**Conflicts of interest**

External funding did not support these post-hoc analyses.

**References**


17. Prien RF, Carpenter LL, Kupper DJ. The definition and operational criteria for treatment outcome of major depressive disorder. A review of the current research literature. Arch Gen Psychiatry. 1993; 48: 796-800.


