Eur Arch Psychiatry Clin Neurosci (2012) 262:291–304
DOI 10.1007/s00406-011-0263-x

ORIGINAL PAPER

Sensitivity to changes during antidepressant treatment: a comparison of unidimensional subscales of the Inventory of Depressive Symptomatology (IDS-C) and the Hamilton Depression Rating Scale (HAMD) in patients with mild major, minor or subsyndromal depression

Isabella Helmreich · Stefanie Wagner · Roland Mergl · Antje-Kathrin Allgaier · Martin Hautzinger · Verena Henkel · Ulrich Hegerl · André Tadić

Received: 29 April 2011 / Accepted: 13 September 2011 / Published online: 30 September 2011

Abstract In the efficacy evaluation of antidepressant treatments, the total score of the Hamilton Depression Rating Scale (HAMD) is still regarded as the 'gold standard'. We previously had shown that the Inventory of Depressive Symptomatology (IDS) was more sensitive to detect depressive symptom changes than the HAMD17 (Helmreich et al. 2011). Furthermore, studies suggest that the unidimensional subscales of the HAMD, which capture the core depressive symptoms, outperform the full HAMD regarding the detection of antidepressant treatment effects. The aim of the present study was to compare several unidimensional subscales of the HAMD and the IDS regarding their sensitivity to changes in depression symptoms in a sample of patients with mild major, minor or subsyndromal depression (MIND). Biweekly IDS-C28 and HAMD17 data from 287 patients of a 10-week randomised, placebo-controlled trial comparing the effectiveness of sertraline and cognitive–behavioural group therapy in patients with MIND were converted to subscale scores and analysed during the antidepressant treatment course. We investigated sensitivity to depressive change for all scales from assessment-to-assessment, in relation to depression severity level and placebo–verum differences. The subscales performed similarly during the treatment course, with slight advantages for some subscales in detecting treatment effects depending on the treatment modality and on the items included. Most changes in depressive symptomatology were detected by the IDS short scale, but regarding the effect sizes, it performed worse than most subscales. Unidimensional subscales are a time- and cost-saving option in judging drug therapy outcomes, especially in antidepressant treatment efficacy studies. However, subscales do not cover all facets of depression (e.g. atypical symptoms, sleep disturbances), which might be important for comprehensively understanding the nature of the disease depression. Therefore, the cost-to-benefit ratio must be carefully assessed in the decision for using unidimensional subscales.

Keywords Minor depression · Hamilton depression rating scale (HAMD17) · Inventory of depressive symptomatology (IDS-C28) · Unidimensional subscales · Sensitivity to change · Depression severity

Ulrich Hegerl and André Tadić authors contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00406-011-0263-x) contains supplementary material, which is available to authorized users.

I. Helmreich (✉) · S. Wagner · A. Tadić
Department of Psychiatry and Psychotherapy,
University Medical Centre Mainz,
Untere Zahlbacher Strasse 8, 55131 Mainz, Germany
e-mail: helmreich_i@psychiatrie.klinik.uni-mainz.de

R. Mergl · U. Hegerl
Department of Psychiatry, University of Leipzig,
Leipzig, Germany

A.-K. Allgaier
Department of Child and Adolescent Psychiatry,
Ludwig-Maximilians-University Munich, Munich, Germany

M. Hautzinger
Institute of Psychology, Department of Clinical and
Developmental Psychology, University of Tübingen,
Tübingen, Germany

V. Henkel
Department of Psychiatry and Psychotherapy,
Ludwig-Maximilians-University Munich, Munich, Germany

Introduction

The continuous assessment of depression severity is an important and decisive element in the optimal clinical treatment of depression as well as in the evaluation of the efficacy of antidepressant treatments. The Hamilton Depression Rating Scale (HAMD, [23, 24]) is still considered to be the 'gold standard' in clinical practice and in research settings, even though it has been widely criticised for conceptual and psychometric shortcomings, especially its multidimensionality and debated sensitivity to depressive change [2, 8, 32, 41, 67]. Besides the development of modern and psychometrically better constructed rating scales such as, e.g., the Inventory of Depressive Symptomatology (IDS, [54, 55]), many studies postulate the use of unidimensional subscales (e.g. [6, 8, 20, 41, 42]) as primary outcome measure to detect treatment differences. The applied logic is that there is an underlying single dimension of depression severity, which can be validly captured with fewer scale items. The postulated advantage of unidimensional subscales is that they capture core depressive symptoms, i.e., the pure antidepressive effect, while too many items of the HAMD17 measure medication side effects (e.g. sleep, weight and sexual problems, somatic anxiety) or dimensions that are border areas of depression (e.g. hypochondriasis). This decreases the ability of the full HAMD to detect real depressive improvement [6, 41, 63]. Another limitation of the HAMD is the inclusion of items that do not change in relation to the severity of the disorder, e.g., loss of weight or insight [19] and therefore do not contribute meaningfully to the overall score as measure for the depression severity.

Different groups have developed the following HAMD17 subscales, which overcome some of these criticisms, especially the multidimensionality: the Bech melancholia scale (Bech6, [7, 9]), the Evans and colleagues scale (Evans6, [19]), the Maier-Phillipp severity subscale (MP6, [37]), the Toronto scale (Toronto7, [39, 40]), the Santen Subscales (Subscale 1 (Santen7) and 2 [56, 57]), the Gibbons' global depression severity (Gibbons8, [22]), the 10-item HAMD (HAMD10, [14]) and the Absence of Depressive and Anxious Mood scale (ADAM2, [59]). Rush and colleagues [53] have also developed a short form of the IDS-C, the Quick Inventory of Depressive Symptomatology-Clinician Rating (QIDS-C16). The items of the QIDS-C16 measure the nine DSM-IV diagnostic criteria for Major Depression (MD), but without any associated symptoms (like anxiety, atypical or melancholic symptoms). The 16 separate items are converted by a scoring system (the highest rated item of each domain contributes to the total score) into the nine DSM-IV symptom criterion domains (see Table 1).

Table 1 Items included in the IDS-C30 and the QIDS16

Item IDS-C30	QIDS16	DSM-IV criterion
1) Sleep onset insomnia	x	Sleep disturbance
2) Mid-nocturnal insomnia	x	
3) Morning insomnia	x	
4) Hypersomnia	x	
5) Mood (sad)	x	Depressive Mood
6) Mood (irritable)		
7) Mood (anxious)		
8) Reactivity of mood		
9) Mood variation		
10) Quality of mood		
11) Appetite (decrease)	x	Weight/appetite change
12) Appetite (increase)	x	
13) Weight (decrease)	x	
14) Weight (increase)	x	
15) Concentration/decision-making	x	Concentration/problems in decision-making
16) Outlook (self)	x	Self-esteem/self-blame
17) Outlook (future)		
18) Suicidal ideation	x	Suicidal ideation
19) Involvement	x	Loss of interest/pleasure
20) Energy/fatigability	x	Energy/fatigue
21) Pleasure/enjoyment (excluding sex)		
22) Sexual interest		
23) Psychomotor slowing	x	Psychomotor agitation/slowing
24) Psychomotor agitation	x	
25) Somatic complaint		
26) Sympathetic arousal		
27) Panic/phobic symptoms		
28) Gastrointestinal		
29) Interpersonal sensitivity		
30) Leaden paralysis/physical energy		
Total No. of Items	16	9
Total Score Range	0–27	

DSM-IV Diagnostic and Statistical Manual of Mental Disorders—Fourth Edition, *IDS-C30* 30-item Inventory of Depressive Symptomatology-Clinician Rating, *QIDS-C16* Quick Inventory of Depressive Symptomatology-Clinician Rating

Table 2 gives an overview of the most commonly used subscales of the HAMD17 and their respective items (consisting of 2 up to 8 items), which contribute to the total scale score. The different scales show a considerable item overlap and include the DSM-IV [1] core symptoms of depression: all subscales (except ADAM2) include item 1 (depressed mood), 2 (guilt), 7 (work and interest) and 10 (anxiety psychic). None of these scales embodies all DSM-IV diagnostic depression criteria as the QIDS-C16 does

Table 2 Items included in shortened forms of the HAMD17

Item HAMD17	Bech6	Evans6	MP6	Toronto7	Santen7	Gibbons8	ADAM2
1. Depressed mood	x	x	x	x	x	x	x
2. Low self-esteem, feelings of guilt	x	x	x	x	x	x	
3. Suicide				x	x	x	
4. Insomnia early (initial insomnia)							
5. Insomnia middle							
6. Insomnia late (terminal insomnia)							
7. Work and activities	x	x	x	x	x	x	
8. Retardation	x		x		x		
9. Agitation			x			x	
10. Anxiety, psychic	x	x	x	x	x	x	x
11. Anxiety, somatic		x		x		x	
12. Somatic symptoms gastrointestinal							
13. Somatic symptoms general (energy)	x	x		x	x		
14. Genital symptoms						x	
15. Hypochondriasis							
16. Loss of weight within the last week							
17. Insight							
Total No. of Items	6	6	6	7	7	8	2
Total score range	0–22	0–22	0–24	0–26	0–26	0–30	0–8

HAMD17 17-item version of the Hamilton Depression Rating Scale, *Bech6* Bech melancholia scale, *Evans6* Evans and colleagues scale, *MP6* Maier-Phillipp severity subscale, *Toronto7* Toronto scale, *Santen7* Santen Subscales 1, *Gibbons8* Gibbons' global depression severity, *ADAM2* Absence of Depressive and Anxious Mood scale

(e.g. loss of appetite/weight or sleep disturbances are missing in all subscales).

Different studies in depressive patients have shown that the above-mentioned subscales are unidimensional and have comparable reliability and equal or even enhanced sensitivity to change compared to the HAMD17 (e.g., [6, 11, 13, 18, 20, 45, 51, 52]).

The Bech6 is the most frequently researched subscale in different populations of depressed patients, e.g., in clinical trials with different antidepressants in inpatients [4, 11–13, 30, 34, 41] or out-patients [16, 45], in combination with psychotherapy [50] or in a naturalistic design in depressed in-patients [36, 38]. A small number of studies have compared several HAMD17 subscales. Ballesteros and colleagues [3] compared the performance of the HAMD17, the Bech6, the Evans6, the MP6, the Toronto7 and the Gibbons8 subscales on a sample of depressive outpatients receiving treatment by standard clinical management in a local multicenter study. The subscales performed equally on sensitivity to change and discriminative power to define remission. Two meta-analyses [18, 20] compared the Bech6, the MP6 and the Gibbons8 subscales: Faries and colleagues [20] demonstrated in their meta-analysis ($n = 2{,}899$) that these subscales slightly outperformed the HAMD17 regarding effect sizes and in detecting antidepressant drug effects in randomised clinical trials (RCT). The Bech6 and the MP6 reached the highest effect sizes: HAMD17 (0.37/0.25), Gibbons8 (0.40/0.27), Bech6 (0.44/0.31), MP6 (0.45/0.32) for the comparison of fluoxetine/tricyclic versus placebos. They also concluded that the use of unidimensional subscales as primary outcome measure could lead to sample size reductions of about one-third as compared to the application of the HAMD17 without losing any statistical power. In a similar reanalysis, Entsuah and colleagues [18] could show for placebo-controlled dose–response trials (fluoxetine/venlafaxine vs. placebo) in patients with MD that the subscales obtained effect sizes 16–76% larger than the HAMD17, indicating a higher sensitivity of the subscales. Again, the Bech6 and the MP6 reached the highest effect sizes. Item 1 (depressed mood) and 10 (anxiety psychic) showed enhanced sensitivity for detecting improvement after treatment, which inspired Silverstone and colleagues [59] to analyse these two items as very brief screening instrument (ADAM2). They found it sufficient for determining changes in the intensity of depression and to predict treatment outcome. Santen and colleagues [57] constructed two subscales with high sensitivity to drug effects and compared them to the Bech6, the MP6 and the HAMD17 [56, 57]. They demonstrated that the subscales have been more sensitive to treatment effects than the more time-consuming HAMD17, although no subscale consistently outperformed the others.

The QIDS-C16—especially in comparative assessment with the HAMD17 and its subscales—is less researched. Trivedi and colleagues [61] as well as Rush and colleagues [51, 52] demonstrated on depressed out-patient samples that the QIDS-C16 performed comparable to or even better than the IDS-C30 or the HAMD17 regarding sensitivity to symptom change and detection of response and depression remission.

To the best of our knowledge, the HAMD subscales and the QIDS-C16 have not yet been compared simultaneously with respect to their sensitivity to detect treatment effects in a population of patients with MIND (mild major, minor or subsyndromal depression). In a previous analysis of the same data set, we showed that the IDS-C28 was more sensitive to detect depressive symptom changes than the HAMD17 [29]. This paper extends previous reports on the comparative assessment of different unidimensional HAMD subscales in different treatment modalities and also relates the results to the performance of the IDS subscale, the QIDS-C16, in a quite representative sample of depressed primary-care patients due to the broad inclusion criteria. Most studies comparing different subscales have been conducted within the framework of RCTs in depressed patients treated with psychopharmacotherapy (e.g., [11, 12, 20, 51, 56, 57]), where patients are usually not representative due to the rigorous exclusion criteria.

The aim of this study was to investigate in a representative population of patients with MIND during the course of an antidepressive treatment: 1) which subscale (Bech6, Evans6, MP6, Toronto7, Santen7, Gibbons8, ADAM2 and QIDS16) is most sensitive in detecting changes in depression severity; 2) whether the subscales differ in their sensitivity to change dependent on the depression severity level; and 3) which subscale best detects placebo–verum differences in different treatment modalities.

Methods

Sample and data collection

The sample consisted of 368 patients with mild major, minor and subsyndromal depression [28]. The data were collected during a prospective, single-centre, 10-week RCT with five treatment arms (sertraline (flexible dosages up to 200 mg/d), pill placebo, manual-guided cognitive–behavioural group therapy (CBT, one individual session and 9 weekly group sessions at 90 min each), guided self-help group (GSG, psychotherapy control condition) and patients' choice condition (i.e. free choice of sertraline or CBT)) testing the effectiveness of sertraline and cognitive–behavioural therapy. It was performed within the framework of the German Research Network on Depression and Suicidality. All participants gave their written informed consent to participate in the study. The study was approved by an independent Ethics Review Committee (Medical Faculty, Ludwig-Maximilians-University Munich, Munich, Germany).

Study design details have been described elsewhere [28]. In brief, patients were referred by primary-care providers to the study centre. The eligibility criteria for the study were as follows: a minimum age of 18 years, a diagnosis of subthreshold (minor) depression, dysthymia or major depressive disorder (according to DSM-IV criteria) with mild to moderate severity (i.e. a HAMD17 total score $\geq$8 and $\leq$22 at screening). Comorbidities such as somatoform and/or anxiety disorders were allowed in order to create a sample as close as possible to the population found in general practice. Acute suicidality, a diagnosis of brief recurrent depression, bipolar affective disorder, addiction (alcohol, benzodiazepines and illicit drugs), schizophrenia, schizotypal personality disorder, delusional disorder, obsessive–compulsive disorder, severe somatic diseases, current psychotherapeutic or antidepressant treatment constituted exclusion. The diagnoses according to DSM-IV criteria were confirmed using a German computer-administered structured clinical interview for DSM-IV (DIA-X, [64]), which is based on the Composite International Diagnostic Interview [65]. As primary efficacy measure for the depressive symptomatology severity, the IDS-C28 [54] and the HAMD17 [24] were assessed biweekly by trained and blinded clinicians. For the subscale comparison, the items of the full scales were converted in the following subscales: Bech6, Evans6, MP6, Toronto7, Santen7, Gibbons8, ADAM2 and QIDS16 (see Tables 1 and 2).

Statistical analysis

Analyses were carried out with the intent-to-treat sample (i.e. all randomised patients). Patients were excluded if baseline assessments for the IDS-C28 and the HAMD17 were not available (IDS-C28: $n = 17$; HAMD17: $n = 1$), and the HAMD17 baseline total score was <8 ($n = 10$). Only patients who had an assessment including all subscale items for both scales (see Tables 1 and 2) at the respective measurement were included in the analysis. This resulted in the following number of patients for each assessment, which were slightly different from our previous analysis [29]: baseline: $n = 287$, week 2: $n = 260$, week 4: $n = 258$, week 6: $n = 245$, week 8: $n = 213$ and week 10: $n = 243$. Baseline clinical and demographical data were computed.

The comparison of the scales (Bech6, Evans6, MP6, Toronto7, Santen7, Gibbons8, ADAM2, QIDS16) regarding sensitivity in detecting changes in depression severity during the course of an antidepressant treatment was

investigated by calculating mean sum scores (±SD) by the assessment for each scale. Agreement in total scores of the QIDS-C16 with the IDS-C28 and the HAMD subscales with the HAMD17 was tested with Pearson's correlations. In order to characterise changes in depressive symptomatology (1) between baseline and each of the five subsequent assessments and (2) from assessment-to-assessment, we calculated the mean sum score changes (Student's t test for dependent samples, $P \leq 0.05$) and analysed effect sizes (di). In order to take between-patient variability into account, we computed an extension of d to individual cases (di) according to the formula from Vittengl et al. [62], i.e., we determined the standardised difference between dependent means by dividing the difference in total scale scores at different time points for each individual patient by the SD of the difference scores. The mean of the resulting di scores is d, which has a SD of 1.0. D was tested for differences between the scales for each assessment by using repeated-measures analyses of variance (ANOVAs) ($P \leq 0.05$). In case the sphericity assumption was not met, the Huynh–Feldt correction was applied. Following post hoc comparisons were performed using the Bonferroni adjustment for multiple comparisons ($P \leq 0.05$). In order to investigate whether the scales indicate the same direction of score change, three groups were formed according to the direction of change: decrease (value = 1), increase (=−1) or no change (=0) in total scores. The direction of change was calculated from assessment-to-assessment (baseline up to week 10) by restructuring the data into cases, resulting in 1,143 data points (subjects x visits) for which concurrent ratings for the IDS and the HAMD were available. Percentage of change in the total group as well as change agreement was analysed by using kappa (κ) statistics.

In order to test whether the sensitivity of the scales is dependent from the depression severity level, three depression severity categories were formed applying the cut-off values established by Paykel [47], which differentiate between treatment effects in general practice. In order to facilitate the understanding, we named Paykel's categories according to the guideline of the National Institute for Health and Clinical Excellence (NICE) [43] as follows: HAMD17 score 0–12 = subthreshold, 13–15 = mild, >15 = moderate. Again, the three change directions of increase, decrease and no change were characterised in percentage for each severity level for the scales from assessment-to-assessment (baseline up to week 10) by restructuring the data (see above). Change agreement between the scales was calculated by using kappa (κ) statistics.

The comparison of the scales regarding the sensitivity to placebo–verum differences (baseline-week 10) was investigated by first calculating the mean sum scores (±SD) by the assessment for each scale and treatment group. All patients with a baseline and a week 10 measurement were included ($n = 243$). The treatment outcome of patients randomised to the patients' choice arm of the study did not differ from that of patients' in the sertraline resp. CBT group [28]. For this reason, patients randomised to the patients' choice condition were assigned to the study arm with identical treatment (sertraline or CBT, resp.), resulting in 4 groups to analyse: sertraline ($n = 95$), placebo ($n = 61$), CBT ($n = 58$), GSG ($n = 29$). Baseline subscale sum scores were tested for differences between treatment groups with oneway ANOVAs ($P \leq 0.05$). Next, the mean sum score changes between baseline and week 10 for each subscale in each treatment condition were computed, and effect sizes di (baseline-week 10) were calculated according to the formula from Vittengl et al. [62], By using ANOVAs ($P \leq 0.05$) and applying the Huynh–Feldt correction in case the sphericity assumption was not met, d was tested for differences between the scales in each treatment group. Lastly, the effect size for each placebo–verum comparison (sertraline vs. placebo, CBT vs. GSG) was calculated with the Hedges's g formula in order to correct for small sample sizes [27] by using the raw difference in means and the standard error from the Student's t test of differences between the two groups. The superiority of the verum is indicated by positive effect sizes. A test of significance of the effect sizes (Z test) was performed ($P \leq 0.05$). All analyses were done using either PASW Statistics 18.0 (SPSS Inc., Chicago, Illinois) or Comprehensive Meta-Analysis 2.2 (Biostat Inc., Englewood, New Jersey).

Results

Patients' characteristics

Two hundred and eighty-seven patients (32.8% men, 67.2% women; mean age (±SD) = 46.51 ± 15.03 years) were included in the analysis. The main diagnoses were as follows: Double Depression (43.6%), Major Depression (31.7%), Depressive Disorder NOS (19.2%), Dysthymic Disorder (3.1%) and Subsyndromal Depressive Disorder (2.4%). Almost half of the patients (43.9%) had a psychiatric comorbid diagnosis. The mean baseline sum scores were 16.50 ± 4.34 points on the HAMD17 and 27.41 ± 7.69 points on the IDS-C28,[1] indicating a moderate depressive symptomatology at study entry. The subscale means are presented in Table 3.

[1] The mean baseline sum scores for the HAMD17 and the IDS-C28 differ slightly from the previously presented figures [29] due to differences in sample size.

Table 3 Scale score comparison regarding mean ± standard deviation (SD)

	Baseline ($n = 287$)	Week 2 ($n = 260$)	Week 4 ($n = 258$)	Week 6 ($n = 245$)	Week 8 ($n = 213$)	Week 10 ($n = 243$)
QIDS-C16	11.80 ± 3.48	10.27 ± 3.97	9.52 ± 4.33	8.76 ± 4.60	8.14 ± 4.41	7.90 ± 5.05
Bech6	8.95 ± 2.43	7.57 ± 3.08	6.93 ± 3.38	6.39 ± 3.54	5.92 ± 3.50	5.60 ± 3.96
Evans6	10.02 ± 2.63	8.63 ± 3.39	7.99 ± 3.74	7.49 ± 3.97	6.92 ± 3.89	6.54 ± 4.30
MP6	8.58 ± 2.31	7.26 ± 2.85	6.70 ± 3.22	6.13 ± 3.40	5.69 ± 3.32	5.39 ± 3.68
Toronto7	10.73 ± 2.95	9.15 ± 3.71	8.46 ± 4.09	7.93 ± 4.37	7.28 ± 4.20	6.90 ± 4.69
Santen7	9.65 ± 2.73	8.10 ± 3.41	7.40 ± 3.73	6.83 ± 3.94	6.27 ± 3.82	5.95 ± 4.34
Gibbons8	11.34 ± 3.04	9.75 ± 3.75	9.10 ± 4.18	8.51 ± 4.56	7.87 ± 4.35	7.47 ± 4.78
ADAM2	3.83 ± 1.15	3.28 ± 1.45	2.93 ± 1.60	2.77 ± 1.66	2.58 ± 1.60	2.38 ± 1.80

QIDS-C16 Quick Inventory of Depressive Symptomatology-Clinician Rating, *Bech6* Bech melancholia scale, *Evans6* Evans and colleagues scale, *MP6* Maier-Phillipp severity subscale, *Toronto7* Toronto scale, *Santen7* Santen Subscales 1, *Gibbons8* Gibbons' global depression severity, *ADAM2* Absence of Depressive and Anxious Mood scale

Table 4 Pearson correlation of the QIDS-C16 with the IDS-C28, respectively, the HAMD subscales with the HAMD17 during treatment

	QIDS-C16	Bech6	MP6	Evans6	Toronto7	Santen7	Gibbons8	ADAM2
Baseline	0.87	0.72	0.74	0.81	0.84	0.77	0.85	0.59
Week 2	0.90	0.86	0.86	0.88	0.90	0.87	0.91	0.71
Week 4	0.93	0.90	0.89	0.92	0.93	0.90	0.94	0.81
Week 6	0.94	0.91	0.89	0.94	0.94	0.92	0.94	0.81
Week 8	0.93	0.92	0.90	0.94	0.95	0.92	0.95	0.82
Week 10	0.95	0.93	0.92	0.94	0.94	0.93	0.95	0.85

QIDS-C16 Quick Inventory of Depressive Symptomatology-Clinician Rating, *IDS-C28* 28-item Inventory of Depressive Symptomatology-Clinician Rating, *HAMD17* 17-item version of the Hamilton Depression Rating Scale, *Bech6* Bech melancholia scale, *Evans6* Evans and colleagues scale, *MP6* Maier-Phillipp severity subscale, *Toronto7* Toronto scale, *Santen7* Santen Subscales 1, *Gibbons8* Gibbons' global depression severity, *ADAM2* Absence of Depressive and Anxious Mood scale

Sensitivity to change during the course of antidepressant treatment

The strength of agreement between sum scores of the QIDS-C16 and the IDS-C28 resp. the HAMD17 and its subscales was excellent for each assessment. The QIDS-C16 yielded the highest correlations ($r \geq 0.87$) followed by the Gibbons8 ($r \geq 0.85$), the ADAM2 the lowest ($r \geq 0.59$, see Table 4). The scales had a similar pattern of mean raw score changes during the study period (see Table 3). Aside from the ADAM2, the Bech6, the MP6 and the Santen7 subscales had the lowest mean sum scores as well as SD on all assessments, i.e., between-patient variability was lowest for these scales.

In order to characterise the changes between the baseline and each of the five subsequent assessments, we analysed the mean sum score changes and determined effect sizes (see Table 5).

Compared to baseline, the mean sum scores decreased significantly at each measurement occasion in all scales ($P < 0.01$ for all t tests). Effect sizes for the subscales were medium to large for each assessment (see Table 5). The effect sizes of the subscales differed significantly for the following assessments: baseline-week 6 ($F(3.73, 909.94) = 3.476$, $P = 0.01$) and baseline-week 8 ($F(3.90, 826.37) = 2.880$, $P = 0.02$). Post hoc comparisons revealed larger effect sizes for the Bech6, the MP6 and the Santen7 subscales on week 6, and for the Bech6 and the Santen7 subscales on week 8 (data not shown).

The analysis of mean sum score changes from assessment-to-assessment showed that most scales (except the ADAM2 between weeks 4 and 6) detected a significant decrease between each of the assessments ($P \leq 0.05$, t tests). Only between weeks 6 and 8, the sum scores did not decline significantly for the QIDS-C16, the Evans6, the Gibbons8 and the ADAM2 ($P \geq 0.05$, t tests). Effect sizes were rather small (see Table 5) and comparable for all subscales ($P \geq 0.05$ for each analysis).

Regarding the direction of change (increase, decrease, no change in total score), the QIDS-C16 identified most change in symptomatology (no change: QIDS-C16 = 12.7% vs. range of other subscales 13.9–30.1%; see Table 6), the ADAM2 the least change (30.1% no change).

The agreement (see Online Resource 1) between the subscales was moderate to large (range κ: 0.40–0.88;

Table 5 Effect sizes (d) for the different scales during the course of treatment and for the treatment subgroups (baseline-week 10)

	QIDS-C16	Bech6	Evans6	MP6	Toronto7	Santen7	Gibbons8	ADAM2
BL-week 2	0.43	0.49	0.46	0.50	0.47	0.50	0.47	0.39
BL-week 4	0.52	0.57	0.54	0.57	0.55	0.58	0.55	0.51
BL-week 6	0.62	0.67	0.59	0.68	0.58	0.64	0.58	0.56
BL-week 8	0.77	0.79	0.73	0.80	0.73	0.79	0.72	0.69
BL-week 10	0.77	0.78	0.75	0.78	0.75	0.78	0.74	0.69
Week 2–4	0.16	0.20	0.18	0.17	0.17	0.19	0.15	0.19
Week 4–6	0.16	0.15	0.14	0.17	0.13	0.15	0.15	0.08
Week 6–8	0.12	0.14	0.13	0.14	0.14	0.15	0.13	0.12
Week 8–10	0.19	0.21	0.23	0.20	0.22	0.20	0.21	0.24
Sertraline	1.03	1.00	0.98	0.99	0.99	1.01	0.94	0.98
Placebo	0.73	0.76	0.66	0.76	0.65	0.73	0.65	0.61
CBT	0.73	0.66	0.71	0.71	0.73	0.69	0.78	0.60
GSG	0.21	0.40	0.31	0.37	0.31	0.39	0.19	0.24

QIDS-C16 Quick Inventory of Depressive Symptomatology-Clinician Rating, *Bech6* Bech melancholia scale, *Evans6* Evans and colleagues scale, *MP6* Maier-Phillipp severity subscale, *Toronto7* Toronto scale, *Santen7* Santen Subscales 1, *Gibbons8* Gibbons' global depression severity, *ADAM2* Absence of Depressive and Anxious Mood scale, *BL* Baseline, *CBT* Cognitive–behavioural group therapy, *GSG* Guided self-help groups

Table 6 Percentage of change (increase, decrease, no change in total score) in the total group and the different depression severity levels (according to HAMD17 cut-off values) for the different scales

	QIDS-C16	Bech6	Evans6	MP6	Toronto7	Santen7	Gibbons8	ADAM2
Total ($n = 1{,}143$)								
Increase	33.9	33.0	33.3	31.7	33.9	32.8	33.7	26.6
No change	12.7	14.9	15.0	16.8	14.2	14.4	13.9	30.1
Decrease	53.5	52.1	51.7	51.5	52.0	52.8	52.4	43.3
Moderate ($n = 432$)								
Increase	25.5	24.3	25.9	23.8	26.4	24.1	25.0	19.7
No change	13.0	13.7	12.0	16.0	11.3	14.1	15.0	30.8
Decrease	61.6	62.0	62.0	60.2	62.3	61.8	60.0	49.5
Mild ($n = 224$)								
Increase	36.6	30.4	29.5	29.9	32.1	31.3	34.8	25.4
No change	8.9	14.7	17.0	15.2	12.5	12.1	8.9	26.3
Decrease	54.5	54.9	53.6	54.9	55.4	56.7	56.3	48.2
Subthreshold ($n = 487$)								
Increase	40.0	41.9	41.7	39.4	41.3	41.3	40.9	33.3
No change	14.2	16.0	16.6	18.3	17.5	15.8	15.2	31.2
Decrease	45.8	42.1	41.7	42.3	41.3	42.9	43.9	35.5

HAMD17 17-item version of the Hamilton Depression Rating Scale, *QIDS-C16* Quick Inventory of Depressive Symptomatology-Clinician Rating, *Bech6* Bech melancholia scale, *Evans6* Evans and colleagues scale, *MP6* Maier-Phillipp severity subscale, *Toronto7* Toronto scale, *Santen7* Santen Subscales 1, *Gibbons8* Gibbons' global depression severity, *ADAM2* Absence of Depressive and Anxious Mood scale

mean ± SD = 0.57 ± 0.16), except for ADAM2 subscale, which had the lowest agreement with the other scales (range κ: 0.25–0.48; mean ± SD = 0.42 ± 0.08). The highest agreements had the subscales that differed only by one additional item, i.e., the Bech6 and the Santen7 subscales as well as the Evans6 and the Toronto7 subscales (both with $\kappa = 0.88$). For all scales, disagreement was highest in the category 'no change' (data not shown).

Sensitivity to change in relation to depression severity

Overall the direction of change ratings (increase, decrease, no change in total score) was similar for the scales for the different severity levels (see Table 6). Most change (decrease or increase) took place in the moderately depressed sample. Most change was indicated by the QIDS-C16 and the Gibbons8 scale (up to 91%). Again, the ADAM2 identified the least change in all depression severity levels (range no change 26–31%), followed by the MP6 subscale (range no change 15–18%). Overall, the agreement in change in total scores between scales (see Online Resource 1) was best for the subthreshold depression level (range $\kappa = 0.22$–0.94; mean $\pm$ SD $= 0.59 \pm 0.18$), followed by the mild (range $\kappa = 0.27$–0.87; mean $\pm$ SD $= 0.54 \pm 0.16$) and the moderate level (range $\kappa = 0.25$–0.81; mean $\pm$ SD $= 0.53 \pm 0.14$). The highest agreements had the subscales that differed only by one additional item, i.e., the Bech6 and the Santen7 subscales as well as the Evans6 and the Toronto7 subscales (range $\kappa = 0.80$–0.93; mean $\pm$ SD $= 0.53 \pm 0.14$). The ADAM2 subscale had the lowest agreement with the other scales, especially in the moderate level (range $\kappa = 0.25$–0.42; mean $\pm$ SD $= 0.39 \pm 0.06$). In all severity levels, disagreement between scales was highest for the category 'no change' (data not shown).

Sensitivity to detect placebo–verum differences

Online Resource 2 shows the results for the mean raw score changes for each subscale during the study period for the different treatment groups. The baseline mean sum scores in the different treatment groups were comparable for each subscale ($P \geq 0.05$ for each analysis). Overall, between-patient variability was again lowest for the ADAM2, the Bech6, the MP6 and the Santen7 subscales. Subscale effect sizes were highest in the sertraline group (range $d = 0.94$–1.03) and lowest in the GSG group (range $d = 0.19$–0.40; see Table 5). However, in each treatment group effect sizes did not differ significantly between the subscales ($P \geq 0.05$ for each analysis). The verum–placebo comparisons are presented in Fig. 1. Effect sizes were small to medium, with higher values for the CBT vs. GSG comparison. In the sertraline vs. placebo comparison, effect sizes were highest for the Santen7 (0.47), the Bech6 (0.46) and the ADAM2 (0.46) subscale, lowest for the QIDS-C16 (0.28). All subscales—except the QIDS-C16—showed that the sertraline treatment was significantly superior to placebo ($P \leq 0.05$ for each analysis). In the CBT vs. GSG group, the Gibbons8 (0.75), the Toronto7 (0.59) and the QIDS-C16 (0.57) yielded the highest values, the Bech6 (0.35), the ADAM2 (0.39), the Santen7 (0.40) and the MP6 (0.43) the lowest. For these scales, sertraline and placebo

Table 7 HAMD17 and QIDS-C16 item effect size analysis (baseline-week 10) for the different treatment groups

HAMD17 item	Treatment	Effect size	QIDS-C16 Item/DSM-IV criterion	Effect size
1. Depressed mood	Sertraline	0.89	5. Mood (sad)	0.85
	Placebo	0.60		0.61
	CBT	0.67		0.45
	GSG	0.27		0.22
2. Low self-esteem, feelings of guilt	Sertraline	0.54	16. Outlook (self)	0.62
	Placebo	0.46		0.48
	CBT	0.50		0.35
	GSG	0.18		0.08
3. Suicide	Sertraline	0.50	18. Suicidal ideation	0.51
	Placebo	0.19		0.16
	CBT	0.51		0.49
	GSG	0.12		0.12
7. Work and activities	Sertraline	0.84	19. Involvement	0.63
	Placebo	0.53		0.47
	CBT	0.42		0.47
	GSG	0.22		0.19
8. Retardation	Sertraline	0.39	23. Psychomotor slowing	0.42
	Placebo	0.19		0.25
	CBT	0.25		0.25
	GSG	0.29		0.20
9. Agitation	Sertraline	0.18	24. Psychomotor agitation	0.10
	Placebo	0.29		0.22
	CBT	0.27		0.14
	GSG	0.09		0.06
13. Somatic symptoms general (energy)	Sertraline	0.53	20. Energy/fatigability	0.69
	Placebo	0.30		0.53
	CBT	0.31		0.65
	GSG	0.38		0.15
10. Anxiety, psychic	Sertraline	0.78		
	Placebo	0.48		
	CBT	0.45		
	GSG	0.17		
11. Anxiety, somatic	Sertraline	0.30		
	Placebo	0.16		
	CBT	0.71		
	GSG	0.11		
14. Genital symptoms	Sertraline	0.24		
	Placebo	0.21		
	CBT	0.34		
	GSG	0.21		
	Sertraline		1.-4. Sleep disturbance	0.30
	Placebo			0.34
	CBT			0.38
	GSG			0.08
	Sertraline		11.-14. Weight/appetite change	0.34
	Placebo			0.25
	CBT			0.23
	GSG			0.04
	Sertraline		15. Concentration/ decision-making	0.64
	Placebo			0.49
	CBT			0.45
	GSG			0.13

HAMD17 17-item version of the Hamilton Depression Rating Scale, *QIDS-C16* Quick Inventory of Depressive Symptomatology-Clinician Rating, *DSM-IV* Diagnostic and Statistical Manual of Mental Disorders—Fourth Edition, *CBT* Cognitive–behavioural group therapy, *GSG* Guided self-help groups

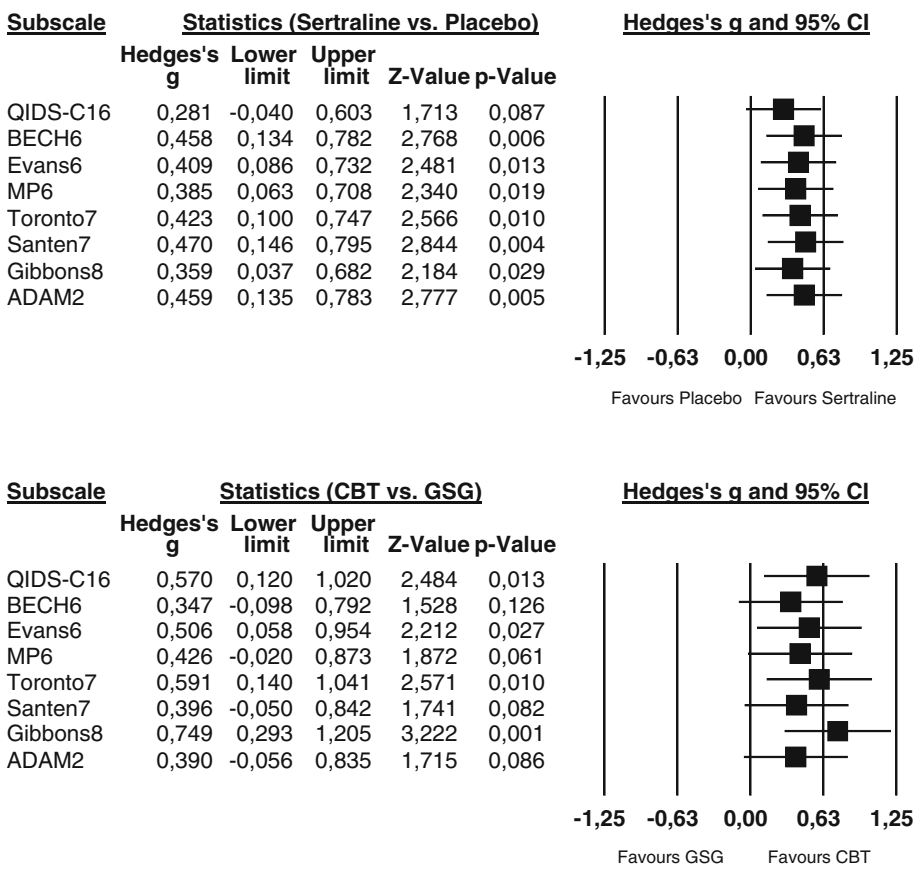

Fig. 1 Placebo–verum comparison (baseline-week 10) for the different subscales. *CI* Confidence interval, *QIDS-C16* Quick Inventory of Depressive Symptomatology-Clinician Rating, *Bech6* Bech melancholia scale, *Evans6* Evans and colleagues scale, *MP6* Maier-Phillipp severity subscale, *Toronto7* Toronto scale, *Santen7* Santen Subscales 1, *Gibbons8* Gibbons' global depression severity, *ADAM2* Absence of Depressive and Anxious Mood scale, *CBT* Cognitive–behavioural group therapy, *GSG* Guided self-help groups

also did not differ significantly in changes from baseline ($P \geq 0.05$ for each analysis).

Discussion

This report compares the unidimensional subscales of the HAMD and the IDS (i.e. the QIDS-C16) regarding their sensitivity to change both during the course of an antidepressant treatment and for different depression severity levels in a representative sample of patients with MIND.

Overall, correspondence in total scores between the IDS-C28, the HAMD17 and their subscales was high ($r \geq 0.72$, except ADAM2 with $r \geq 0.59$) and comparable to values found in other studies among depressive patients (e.g., [5, 50, 51, 54]). The course of mean symptom change was similar for all subscales: depressive symptoms decreased rapidly in both scales in the early course of treatment (i.e. in the first 2 weeks) and were more stable afterwards. The Bech6, the MP6, the Santen7 and the Toronto7 subscales were slightly more sensitive in detecting a significant change between weeks 6 and 8. This indicates that these scales are better able to identify the small changes in symptomatology in this stage of the treatment. Regarding effect sizes, the scales also differed. Overall, effect sizes increased over time and were moderate to large for all scales compared to baseline, small from assessment-to-assessment. The Bech6, the MP6 and the Santen7 subscales showed an advantage in effect sizes between baseline and week 6, the Bech6 and the Santen7 subscales in week 8. These scales also had the lowest scale score means and between-patient variability on all assessments, indicating that the individual data points are close to the mean. The chosen items seem to measure the proposed underlying single dimension of depression severity quite well for patients with MIND.

Compared to other studies among patients with moderate to severe MDD (e.g., [3, 50, 51, 62]), effect sizes were lower, likely due to the lower depression severity scores of the MIND population, but comparable to the study of patients with MIND (e.g., [50]). In our study, only the Bech6, the MP6 and the Santen7 subscales were more powerful to observe treatment effects compared to the full scales IDS-C28 and the HAMD17, while the QIDS-C16 did not perform better than the full scales [29].

The subscales also differed in their ability to detect change in symptomatology (increase or decrease), for the total sample as well as for the different severity levels (subthreshold, mild, moderate), depending on the items included in the rating scale. In the lower depression

severity level, the QIDS-C16 performed better than the other scales, due to its wider item range, which also covers cognitive symptoms, which are common among patients with mild depression [31, 48]. In the moderate depression level, the subscales that included items on somatic symptoms (i.e. the Bech6, the Evans6, the Toronto7 and the Santen7 subscale) detected the most change. These are symptoms that are more common among severely depressed patients [19]. The more items each subscale included, the better it detected changes, i.e., the ADAM2 performed worst. Overall, the agreement in change between the scales was moderate to large, with slightly higher correlations in the lower depression levels. As described by different authors (e.g., [58]), this might be a statistical effect due to a greater homogeneity of the sample, i.e., weaker correlations are a result of an increase in total score range and SD (e.g. mean change HAMD17 ± SD (range): moderate depression level = 0.40 ± 0.87 (0.31–0.48)/mild: 0.28 ± 0.91 (0.16–0.40)/subthreshold: −0.04 ± 0.93 (−0.13–0.04)). Change agreement was also better for scales that had similar items included.

Analyses of effect size scores (baseline-week 10) for the different treatment groups showed that the effect sizes were comparable between the scales. They were highest in the sertraline group (range $d = 0.94$–1.03) and lowest in the GSG (range $d = 0.19$–0.40) group. The placebo (range $d = 0.61$–0.76) and CBT (range $d = 0.60$–0.78) group had similar effect sizes. A post hoc analyses of the effect sizes (baseline-W10) for each item (see Table 7) revealed that most items detected change very well in the sertraline group (e.g. 'depressed mood' d sertraline = 0.89, 'work and activities' d sertraline = 0.84), but were less sensitive in the other groups (e.g. 'depressed mood' d placebo = 0.60, d CBT = 0.67, d GSG = 0.27). The item 'somatic anxiety' was the only item that had the highest impact on the CBT group (d CBT = 0.71 vs. d other groups < 0.30). Subscales that included this item (i.e. the Evans6, Toronto7 and Gibbons8) had also higher effect sizes in the CBT compared to the placebo group (see Table 7). The data on the CBT group support the non-specific treatment effect of CBT (as well as placebo), i.e., improvement occurs in various different depressive and anxiety symptoms [33]. Regarding the sensitivity to detect placebo–verum differences, all subscales (except the QIDS-C16) detected a significant treatment effect in the sertraline–placebo comparison, with best values for the Santen7, the Bech6 and the ADAM2 subscale. Effect sizes were comparable to other studies among patients with MD treated with a selective serotonin reuptake inhibitor (SSRI; e.g., [18, 20]). The post hoc analysis of the scale item effect sizes (see Table 7) revealed that these subscales included the HAMD17 items 'depressed mood', 'suicide', 'work and activities', 'psychic anxiety' and 'somatic symptoms general', which were most sensitive to the difference in improvement in depression severity between sertraline and placebo. The missing sensitivity of the QIDS-C16 can be explained by the higher number of items, i.e., the subscale item are more heterogeneous regarding their treatment sensitivity resulting in a lower sensitivity (see Table 7). The effect sizes for the CBT vs. GSG comparison were higher than for the sertraline vs. placebo comparison; for the Gibbons8, the Toronto7, the QIDS-C16 and the Evans6 subscale even clinically relevant according to the NICE standards [43] (i.e. an effect size of 0.50 is considered to be clinically relevant). The HAMD17 subscales all included the item 'somatic anxiety', which was most distinctive between CBT and GSG, and did not include the items 'retardation' and 'somatic symptoms general', which showed the smallest effect between CBT and GSG. The items of the QIDS-C16 were all quite well able to differentiate between CBT and GSG (see Table 7). The large difference in effect sizes between the sertraline vs. pill placebo (small to medium) and the CBT vs. GSG (medium to large) comparison is a result of the worse treatment outcome of the GSG group (the worst of all groups). The reasons for this (e.g. the 'nocebo' effect, the inability to blind therapists and patients) are discussed in more detail elsewhere (see [28]) and illustrate the difficulties to invent an adequate psychotherapy 'placebo' condition.

The main study limitations come from the non-independence of the subscales, because we used the collected data from the IDS-C28 and HAMD17 to derive the different subscale scores. Therefore, independent comparisons among the scales are not possible and our results might be inflated. However, the bias should be the same for each scale; therefore, a comparison between the scales is justifiable. Additionally, treatment side effects (e.g. agitation) could have affected the subscale scores differently in the total as well as in the treatment groups. Treatment effects could have been reduced in subscales including items like agitation (e.g. in the MP6) and also through the different incidence of side effects: 10% in the sertraline group withdrew due to the adverse events (e.g. tremor, nausea, agitation), 4% in the placebo group, 2% in the CBT group and 6% in the patients' choice sertraline arm (for details see [28]). Another limitation is the same rater completed both interviews (IDS-C28 and HAMD17), thus the interviewer was not blinded to the results of the other scale, which could have enhanced the agreement between scales. However, by having the same rater judging the symptomatology of one patient, existing differences between the scales are not affected by the inter-rater variability. Rater agreement in our study was high with 95% [28], but it still has an influence on the score ratings with 5% rater disagreement. Because independent ratings also have a negative impact on study time and costs, the advantages of

having one rater for both scales outweighed the disadvantages. However, an independent comparison of the subscales by using different randomisation arms as McIntyre and colleagues [40] have applied it in their validation of the Toronto7 scale would be interesting and warranted. The sample sizes for the CBT and GSG arm were rather low, due to an early stopping of CBT and GSG after planned interim analyses (for details, see [28]). Therefore, an analysis of depression subgroups in the placebo–verum comparisons was not possible. Further studies with higher sample sizes, especially for the psychotherapy treatment condition, are needed in order to substantiate our findings in the total group as well as for different depression severity levels. Another study limitation is that the real depression severity change cannot be validated by an external criterion. However, the full scales (IDS-C28 and HAMD17) had a high concurrent validity (Pearson correlations over 0.85), indicating that the scales accurately measure depression. The lack of an external validation criterion also means that it cannot be fully excluded that the higher change rates detected by the QIDS-C16 could also be a result of a higher false positive rate, i.e., the QIDS-C16 could detect changes in symptomatology where in fact there is no change, while the HAMD subscales would produce less false positives, e.g., that it does not change when there is no change. An additional source of a potentially higher variance of the QIDS-C is the fact that it has more items than the HAMD subscales. Therefore, further studies should focus on the validation of our results by using an external validation criterion like an independent global expert severity assessment (LEAD approach) involving all available data (e.g. clinical, biological, physiological assessments) [35].

In summary, the subscales (except the ADAM2, which performed worst in all domains) have been well able to measure depressive symptomatology in a quite representative sample of depressed primary-care patients with mild major, minor or subsyndromal depression. In our study, especially the Bech6, the MP6 and the Santen7 subscales were more powerful to observe treatment effects. Regarding the effect sizes, the QIDS-C16 performed worse than most subscales, but detected more changes, due to its wider item range. The sensitivity to detect change in symptomatology depended on the items included in the rating scales: in the moderately depressed sample subscales embodying items which are common among severely depressed patients (e.g. somatic symptoms) performed slightly better and in the mildly depressed sample subscales which include items associated with mild depression (e.g. cognitive symptoms). One reason for the lack of sensitivity to treatment effects might be that the QIDS-C16 does not include any anxiety item, because it does not belong to the main DSM-IV depression criteria. Nevertheless, anxiety is considered to be an important co-morbid symptom of depression (e.g., [6, 26]). Patients presenting phobic anxiety symptoms report a poorer quality of life [17], take longer to recover, and typically show less response and remission to antidepressants [21, 44, 46, 49, 66]. Studies [9, 10, 56, 57] have demonstrated that the psychic anxiety is a very sensitive and useful item in assessing response to treatment and it is therefore included in all HAMD subscales.

When compared with the performance of the full scales (HAMD17 and IDS-C28, see published data in Helmreich et al. [29]), the full scales were better able to detect small changes in depressive symptomatology than the short scales, due to their wider range of items. Regarding pill placebo–verum differences, the short scales were also more sensitive than the full versions of the respective scales (*d* sertraline vs. placebo HAMD17 = 0.33, IDS-C28 = 0.36). However, for the CBT vs. GSG comparison, the full scales achieved higher effect sizes than most subscales (*d* CBT vs. GSG HAMD17 = 0.72, IDS-C28 = 0.60). Our results indicate that the treatment modality is another factor that has to be taken into consideration when selecting the appropriate scale to measure efficacy. Results of other studies show that different scales might be appropriate for different medications (e.g., [15, 20, 25]). Hamilton and Shapiro [25] could show that the HAMD17 scale seems to be more sensitive to serotonin-acting antidepressants, while the Montgomery Åsberg Rating Scale (MADRS, [42]) was more sensitive to noradrenaline-acting antidepressants due to the different weight of individual items (e.g. decreased appetite). Ruhé and colleagues [50] compared different treatment modalities (pharmacotherapy with a combination of psychopharmacological treatment and Short Psychodynamic Supportive Psychotherapy), but did not find any significant differences. However, they were not able to compare both therapies as stand alone condition. Due to the limited evidence on subscale comparisons in trials including a psychotherapy treatment condition, our results are preliminary and have to be replicated in different depressive populations. The differing sensitivity of the scales could lead to individually adapted measurement-based care approach as, e.g., Trivedi [60] proposes it. The Santen7 subscale, which was specifically developed for measuring drug–placebo differences, seems to be a very interesting option in this field. Compared to the Bech6 and the MP6 subscales it also includes the item suicide, according to Riedel and colleagues [49] a highly significant predictor for treatment response, if it reaches a score value of $\geq$3 points. However, to date there is a dearth of studies validating and evaluating this scale in different depressive populations and treatment conditions.

In conclusion, the short scales performed similarly regarding their sensitivity to change and to detect treatment

differences during the treatment course. However, the Bech6, the MP6 and the Santen7 subscales were more powerful to observe treatment effects, especially in the sertraline group, while the Gibbons8 subscale and the full scales were superior in the CBT group. Compared to the full scales, markedly the IDS-C28 outperforms the short scales regarding the sensitivity to detect small changes in depression symptomatology. The Santen7 subscale seems to be the most promising option to use in RCT comparing antidepressants to placebo, because it was constructed to assess drug–placebo differences and it also includes the item suicide, a useful predictor for treatment response. The Gibbons8 subscale and the full scales seem to measure change in the CBT group best. However, these results are preliminary and have to be replicated. Therefore, on the one hand, subscales are a time- and cost-savings option in judging drug therapy outcomes, especially in antidepressant treatment efficacy studies. On the other hand, however, subscales do not cover all facets and symptom ranges of depression, which might be important for comprehensively understanding the nature of the disease depression and in the clinical routine for, e.g., capturing symptom profiles, identifying subtypes, studying biological correlates and phenotypes, or detecting small changes in symptomatology in order to make decisions on the most effective treatment and predicting long-term outcomes. Therefore, the cost-to-benefit ratio must be carefully assessed in the decision for solely using unidimensional subscales.

A highly recommended option—especially for RCTs, which often rely on full scales—would be the inclusion of subscale analyses as well as profile scores (e.g. a score for the pure antidepressive effect, somatic complaints, anxiety, suicide, atypical symptoms, etc.) as secondary outcome criteria. This would lead to a better understanding of the advantages as well as the shortcomings of the unidimensional subscales.

Acknowledgments We are grateful to all patients who participated in the study, the general practitioners who screened the patients within the network of medical practices, Nuremberg North, especially Veit Wambach, MD, Vanadis Kamm-Kohl, MD and the practice staff. We are indebted to Patrick Bussfeld, MD, Ute Hägele, MD, Winfried Scheunemann†, Michael Schütze, MD, the physicians who investigated the patients at the specialised study center, Michaela König, MSc, Stephanie Lösch, MSc, Michael Stürmer, MSc who monitored the progress of the study and Professor Hans-Jürgen Möller, MD for his contributions to design and methods of the RCT reported here (trial registration: clinicaltrials.gov; URL: http://www.clinicaltrials.gov; registration number: NCT00226642). We are also thankful to Suzanne Snead, PhD, for the fruitful discussion of the manuscript.

Funding Funding for the study was provided by the German Ministry for Education and Research (BMBF; grant: 01 GI 9922/0222/0452) within the promotional emphasis 'German Research Network on Depression and Suicidality'. The BMBF had no further role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. American Psychiatric Association (1994) Diagnostic and statistical manual of mental disorders, 4th edn. American Psychiatric Press, Washington DC
2. Bagby M, Ryder AG, Schuller DR, Marshall MB (2004) The Hamilton depression rating scale: has the gold standard become a lead weight? Am J Psychiatry 161:2163–2177
3. Ballesteros J, Bobes J, Bulbena A, Luque A, Dal-Ré R, Ibarra N, Güemes I (2007) Sensitivity to change, discriminative performance, and cutoff criteria to define remission for embedded short scales of the Hamilton depression rating scale (HAMD). J Affect Disord 102:93–99
4. Bech P (2001) Meta-analysis of placebo-controlled trials with mirtazapine using the core items of the Hamilton Depression Scale as evidence of a pure antidepressive effect in the short-term treatment of major depression. Int J Neuropsychopharmacol 4:337–345
5. Bech P (2008) The use of rating scales in affective disorders. Eur Psychiatr Rev 1:14–18
6. Bech P (2010) Is the antidepressive effect of second-generation antidepressants a myth? Psychol Med 40:181–186
7. Bech P, Gram LF, Dein E, Jacobsen O, Vitger J, Bolwig TG (1975) Quantitative rating of depressive states. Acta Psychiatr Scand 51:161–170
8. Bech P, Rafaelsen OJ (1980) The use of rating scales exemplified by a comparison of the Hamilton and the Bech Rafaelsen Melancholia scale. Acta Psychiatr Scand 62(S285):128–131
9. Bech P, Allerup P, Gram LF, Reisby N, Rosenberg R, Jacobsen O, Nagy A (1981) The Hamilton depression scale. Evaluation of objectivity using logistic models. Acta Psychiatr Scand 63:290–299
10. Bech P, Allerup P, Reisby N, Gram LF (1984) Assessment of symptom change from improvement curves on the Hamilton Depression Scale in trials with antidepressants. Psychopharmacology 84:276–281
11. Bech P, Ciadella P, Haugh MC, Hours A, Boissel JP, Birkett MA, Tolllefson GD (2000) Meta-analysis of randomised controlled trials of fluoxetine vs. placebo and tricyclic antidepressants in the short-term treatment of major depression. Br J Psychiatry 176:421–428
12. Bech P, Tanghoj P, Andersen HF, Overo K (2002) 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 163:20–25
13. Bech P, Boyer P, Germain JM, Padmanabhan K, Haudiquet V, Pitrosky B, Tourian KA (2010) HAM-D17 and HAM-D6 sensitivity to change in relation to desvenlafaxine dose and baseline depression severity in major depressive disorder. Pharmacopsychiatry 43:271–276
14. Benazzi F (1998) A 10-item hamilton depression rating scale to measure major depressive episode severity in outpatients. Int J Geriatr Psych 13:568–574

15. Bent-Hansen J, Lunde M, Klysner R, Andersen M, Tanghøj P, Solstad K, Bech P (2003) The validity of the depression rating scales in discriminating between citalopram and placebo in depression recurrence in the maintenance therapy of elderly unipolar patients with major depression. Pharmacopsychiatry 36:313–316
16. Carmody TJ, Rush AJ, Bernstein I, Warden D, Brannan S, Burnham D, Woo A, Trivedi MH (2006) The Montgomery Åsberg and the Hamilton ratings of depression: a comparison of measures. Eur Neuropsychopharmacol 16:601–611
17. Coryell W, Endicott J, Andreasen NC, Keller MB, Clayton PJ, Hirschfield RM, Scheftner WA, Winokur G (1988) Depression and panic attacks: the significance of overlap as reflected in follow-up and family study data. Am J Psychiatry 145:293–300
18. Entsuah R, Shaffer M, Zhang J (2002) A critical examination of the sensitivity of unidimensional subscales derived from the Hamilton Depression Rating Scale to antidepressant drug effects. J Psychiatr Res 36:437–448
19. Evans KR, Sills T, DeBrota DJ, Gelwicks S, Engelhardt N, Santor D (2004) An item response analysis of the Hamilton Depression Rating Scale using shared data from two pharmaceutical companies. J Psychiatr Res 38:275–284
20. Faries D, Herrera J, Rayamajhi J, DeBrota D, Demitrack M, Potter WZ (2000) The responsiveness of the Hamilton Depression Rating Scale. J Psychiatr Res 34:3–10
21. Fava M, Rush AJ, Alpert JE, Balasubramani GK, Wisniewski SR, Carmin CN, Biggs MM, Zisook S, Leuchter A, Howland R, Warden D, Trivedi MH (2008) Difference in treatment outcome in outpatients with anxious versus nonanxious depression: a STAR*D report. Am J Psychiatry 165:342–351
22. Gibbons RD, Clark DC, Kupfer DJ (1993) Exactly what does the Hamilton Depression Rating Scale measure? J Psychiatr Res 27:259–273
23. Hamilton M (1960) A rating scale for depression. J Neurology, Neurosurgery, Psychiatry 23:56–62
24. Hamilton M (1967) Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 6:278–296
25. Hamilton M, Shapiro CM (1990) Depression. In: Peck DF, Shapiro CM (eds) Measuring human problems. Wiley, Chichester, pp 25–65
26. Hecht H, von Zerssen D, Wittchen H-U (1990) Anxiety and depression in a community sample: the influence of comorbidity on social functioning. J Affect Disord 18:137–144
27. Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic Press, New York, pp 78–85
28. Hegerl U, Hautzinger M, Mergl R, Kohnen R, Schütze M, Scheunemann W, Allgaier A-K, Coyne J, Henkel V (2010) Effects of pharmaco- and psychotherapy in depressed primary care patients. A randomized, controlled trial including a patient choice arm. Int J Neuropsychopharmacol 13:31–44
29. Helmreich I, Wagner S, Mergl R, Allgaier A-K, Hautzinger M, Henkel V, Hegerl U, Tadić A (2011) The Inventory of Depressive Symptomatology (IDS-C28) is more sensitive to changes in depressive symptomatology than the Hamilton Depression Rating Scale (HAMD17) in patients with mild major, minor or subsyndromal depression. Eur Arch Psychiatry Clin Neurosci 261:357–367
30. Hooper CL, Bakish D (2000) An examination of the sensitivity of the six-item Hamilton Rating Scale for Depression in a sample of patients suffering from major depressive disorder. J Psychiatry Neurosci 25:178–184
31. Howland RH, Schettler PJ, Rapaport MH, Mischoulon D, Schneider T, Fasiczka A, Delrahiem K, Maddux R, Lightfoot M, Nierenberg AA (2008) Clinical features and functioning of patients with minor depression. Psychother Psychosom 77:384–389
32. Iannuzzo R, Jaeger J, Goldberg J, Kafantaris V, Subletteet M (2006) Development and reliability of the HAM-D/MADRS Interview: An integrated depression symptom rating scale. Psychiat Res 145:21–37
33. Ilardi SS, Craighead WE (1994) The role of nonspecific factors in cognitive-behavior therapy for depression. Clin Psychol Sci and Practice 1:138–156
34. Lecrubier Y, Bech P (2007) The Ham D(6) is more homogenous and as sensitive as the Ham D(17). Eur Psychiatry 22:252–255
35. Maier W (1990) The Hamilton Depression Scale and its alternatives: a comparison of their reliability and validity. In: Bech P, Coppen A (eds) The Hamilton Scales. Springer-Verlag, Berlin, pp 64–71
36. Maier W, Heuser I, Philipp M, Frommberger U, Demuth W (1988) Improving depression severity assessment—II. Content, concurrent and external validity of three observer depression scales. J Psychiatr Res 22:13–19
37. Maier W, Philipp M (1985) Improving the assessment of severity of depressive states: a reduction of the Hamilton depression scale. Pharmacopsychiatry 18:114–115
38. Maier W, Philipp M, Heuser I, Schlegel S, Buller R, Wetzel H (1988) Improving depression severity assessment—I. Reliability, internal validity and sensitivity to change of three observer depression scales. J Psychiatr Res 22:3–12
39. McIntyre R, Kennedy S, Bagby RM, Bakish D (2002) Assessing full remission. J Psychiatry Neurosci 27:235–239
40. McIntyre R, Konarski JZ, Mancini DA, Fulton KA, Parikh SV, Grigoriadis S, Grupp LA, Bakish D, Filteau M-J, Gorman C, Nemeroff CB, Kennedy SH (2005) Measuring the severity of depression and remission in primary care: validation of the HAMD-7 scale. CMAJ 173:1327–1334
41. Möller H-J (2001) Methodological aspects in the assessment of severity of depression by the Hamilton Depression Scale. Eur Arch Psychiatry Clin Neurosci 251(suppl 2):13–20
42. Montgomery SA, Åsberg M (1979) A new depression scale designed to be sensitive to change. Br J Psychiatry 134:382–389
43. National Institute for Health, Clinical Excellence (2009) Depression: the Treatment and Management of Depression in Adults. National Clinical Practice Guideline 90. National Institute for Health and Clinical Excellence, London
44. Nutt D (1997) Management of patients with depression associated with anxiety symptoms. J Clin Psychiatry 58(Suppl 8):11–16
45. O'Sullivan RL, Fava M, Agustin C, Baer L, Rosenbaum JF (1997) Sensitivity of the six-item Hamilton depression rating scale. Acta Psychiatr Scand 95:379–384
46. Papakostas GI, Larsen K (2011) Testing anxious depression as a predictor and moderator of symptom improvement in major depressive disorder during treatment with escitalopram. Eur Arch Psychiatry Clin Neurosci 261:147–156
47. Paykel ES (1990) The use of the Hamilton Depression Scale in general practice. In: Bech P, Coppen A (eds) The Hamilton Scales. Springer-Verlag, Berlin, pp 40–48
48. Rapaport MH, Judd LL, Schettler PJ, Yonders KA, Thase ME, Kupfer DJ, Frank E, Plewes JM, Tollefson GD, Rush AJ (2002) A descriptive analysis of minor depression. Am J Psychiatry 159:637–643
49. Riedel M, Möller H-J, Obermeier M, Adli M, Bauer M, Kronmüller K, Brieger P, Laux G, Bender W, Heuser I, Zeiler J, Gaebel W, Schennach-Wolff R, Henkel V, Seemüller F. (2011) Clinical predictors of response and remission in inpatients with depressive syndromes. J Affect Disord doi:10.1016/j.jad.2011.04.007
50. Ruhé HG, Dekker JJ, Peen J, Holman R, de Jonghe F (2005) 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. Comprehensive Psychiatry 46:417–427
51. Rush AJ, Bernstein IH, Trivedi MH, Carmody TJ, Wisniewski S, Mundt JC, Shores-Wilson K, Biggs MM, Woo A, Nierenberg AA, Fava M (2006) An Evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression: A Sequenced Treatment Alternatives to Relieve Depression Trial Report. Biol Psychiatry 59:493–501
52. Rush AJ, Carmody TJ, Ibrahim HM, Trivedi HM, Biggs MM, Shores-Wilson K, Crismon ML, Toprac MG, Kashner TM (2006) Comparison of self-report and clinician ratings on two inventories of depressive symptomatology. Psychiatr Serv 57:826–837
53. Rush AJ, Carmody TJ, Reimitz PE (2000) The Inventory of Depressive Symptomatology (IDS): Clinician (IDS-C) and self-report (IDS-SR) ratings of depressive symptoms. Int J Methods Psychiatr Res 9:45–59
54. Rush AJ, Gilles DE, Schlesser MA, Fulton CL, Weissenburger J, Burns C (1986) The Inventory for Depressive Symptomatology (IDS): preliminary findings. Psychiatry Res 18:65–87
55. Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi HM (1996) The Inventory of Depressive Symptomatology (IDS): psychometric properties. Psychol Med 26:477–486
56. Santen G, Danhof M, Pasqua OD (2009) Sensitivity of the Montgomery Asberg Depression Rating Scale to response and its consequences for the assessment of efficacy. J Psychiatr Res 43:1049–1056
57. Santen G, Gomeni R, Danhof M, Pasqua OD (2008) Sensitivity of the individual items of the Hamilton depression rating scale to response and its consequences for the assessment of efficacy. J Psychiatr Res 42:1000–1009
58. Senra C, Polaino A (1993) Concordance between clinical and self-report depression scales during the acute phase and after treatment. J Affect Disord 27:13–19
59. Silverstone PH, Entsuah R, Hackett D (2002) Two items on the Hamilton Depression rating scale are elective predictors of remission: comparison of selective serotonin reuptake inhibitors with the combined serotonin/norepinephrine reuptake inhibitor, venlafaxine. Int Clin Psychopharmacol 17:273–280
60. Trivedi MH (2009) Tools and strategies for ongoing assessment of depression: a measurement-based approach to remission. J Clin Psychiatry 70(Suppl 6):26–31
61. Trivedi MH, Rush AJ, Ibrahim HM, Carmody TJ, Biggs MM, Suppes T, Crismon ML, Shores-Wilson K, Toprac MG, Dennehy EB, Witte B, Kashner TM (2004) 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 34:73–82
62. Vittengl JR, Clark LA, Kraft D, Jarrett R (2005) Multiple measures, methods, and moments: a factor-analytic investigation of change in depressive symptoms during acute-phase cognitive therapy for depression. Psychol Med 35:693–704
63. Walczak DD, Apter JT, Halikas JA, Borison RL, Carman JS, Post GL, Patrick R, Cohn JB, Cunningham LA, Rittberg B, Preskorn SH, Kang JS, Wilcox CS (1996) The oral dose-effect relationship for fluvoxamine: a fixed-dose comparison against placebo in depressed outpatients. Ann Clin Psychiatry 8:139–151
64. Wittchen H-U, Pfister H (eds) (1997) DIA-X-Interview, Instruktionsmanual zur Durchführung von DIA-X Interviews. Swets & Zeitlinger, Frankfurt am Main
65. WorldHealth Organization (WHO) (1993) Composite International Diagnostic Interview, Version 1.1. World Health Organization, Geneva
66. Yang H, Chuzi S, Sinicropi-Yao L, Johnson D, Chen Y, Clain A, Baer L, McGrath PJ, Stewart JW, Fava M, Papakostas GI (2010) Type of residual symptom and risk of relapse during the continuation/maintenance phase treatment of major depressive disorder with the selective serotonin reuptake inhibitor fluoxetine. Eur Arch Psychiatry Clin Neurosci 260:145–150
67. Zimmerman M, Posternak MA, Marshall MB (2005) Is it time to replace the Hamilton Depression Rating Scale as the primary outcome measure in treatment studies of depression? J Clin Psychopharmacol 25:105–110