Eur Arch Psychiatry Clin Neurosci (2011) 261:37–46
DOI 10.1007/s00406-010-0106-1

ORIGINAL PAPER

Validity of the Definite and Semidefinite Questionnaire version of the Hamilton Depression Scale, the Hamilton Subscale and the Melancholia Scale. Part I

Jesper Bent-Hansen · Per Bech

Received: 4 June 2009 / Accepted: 3 February 2010 / Published online: 2 March 2010

Abstract Instruments for self-rating in depression are available, but their psychometric properties have not been fully explored; discrepancies with clinician ratings have been identified. This study was longitudinal with 85 patients fulfilling the DSM-III-R diagnosis of Seasonal Affective Disorder. Self-reporting versions (definitely and semidefinitely anchored) corresponding to the Hamilton Depression Scale (HAMD), the Hamilton Subscale (HAM_6), and the Bech–Rafaelsen Melancholia Scale (MES) were compared to each other and the clinician-rated version. The unidimensional property of the sum score in each scale was tested by the item-response theory model ad modum Rasch. The scales were also tested for their sensitivity to discriminate between placebo and citalopram therapy. The sum scores and the sum score variances of the definite self-rating versions did not differ significantly from the sum scores of the corresponding observer scales at any of the five time points. The semidefinite scales significantly over-scored at all time points. The convergent validity between corresponding definite self-ratings and observer ratings was very high with correlations exceeding 0.90. Only item responses from the MES, the HAM_6, and their corresponding definite versions of the self-rating questionnaires DMQ and $DHAM_6$ were accepted by the Rasch analysis, and only these four valid scales discriminated significantly between the effect of citalopram and placebo treatment. Our results are limited to patients with moderate depression. Two new self-report scales with unparalleled construct validity, reliability, sensitivity, and convergent validity have been identified (DMQ and $DHAM_6$). We have also identified a crucial importance of format for the means and variances of self-rating scales. These findings are of high practical and scientific value.

Keywords Analogous observer and self-reported rating scales · Definitely vs. Semidefinitely anchored self-rated items · IRT (Rasch) validation of self-reported scales · Variance of self-rated scales

J. Bent-Hansen (✉)
Department of Psychiatric Research, Department of Psychiatry,
Frederiksberg Hospital Department,
2000 Frederiksberg, Denmark
e-mail: jbent-hansen@dadlnet.dk

P. Bech
Psychiatric Research Unit, Frederiksborg General Hospital,
3400 Hillerød, Denmark

Introduction

Reliable and valid self-rating scales covering the components of established observer-rating scales for depression have great potential usefulness [13]. With modern electronic technology, such self-rating scales would permit large-scale investigations at low cost, as a major advantage of self-rating scales over observer scales is their reduced professional time consumption. In epidemiological research papers that rely only on self-assessed data are being published more frequently [1].

Another major advantage of self-rating scales is that they provide a degree of standardization that can be difficult to achieve with observer ratings, which may be influenced by individual raters.

Systematic over-scoring of self-reported ratings in comparison to observer ratings has been found to be a major difference between the two [16, 33, 36]. This discrepancy has mainly been considered a patient bias and little attention has therefore been paid to the psychometric aspects of the self-rating scales. These aspects can be more readily investigated if the self-rating scales are analogous versions of the observer scales [11, 14]. The self-rated scales can be compared directly with the analogous observer scales, testing the assumption (null hypothesis) that they do not differ. The item content reflecting the concept of depression and the *structure* of self-rated scales may be decisive factors for the degree of association with clinician-rated scales [11, 15]. Other important factors are the easy readability and understanding of the item-response options of the questionnaire. Intensity ratings are preferable to frequency ratings of depression symptoms [19].

Semi-anchored self-rating versions (in which the adverbs or adjectives used lack precise semantic definitions) of the Hamilton Depression Scale [17, 18] and the Bech–Rafaelsen Melancholia Scale (MES) [8] have previously been found to over-score significantly as opposed to more anchored versions (in which the behavior corresponding to each level of severity is precisely described)—the Definite Hamilton Questionnaire (DHQ) and the Definite Melancholia Questionnaire (DMQ) [11] which substantially agreed with the observer-rated scales. The structure of the questionnaires and the definitions of the terms “definite” (anchored) and “semidefinite” (semi-anchored) have been described in detail elsewhere [11]. Most importantly, the definite questionnaires are in close accordance with the item-definition manual accepted by Hamilton himself [6] (Fig. 1).

The patients in the Bent-Hansen et al. [11] study fulfilled the criteria of major depressive disorder (DSM-IIIR). We investigated to what extent their findings of a close association between definite self-rating scales and the corresponding observer scales and of the over-scoring of the semidefinite questionnaires [11] were applicable to other types of depressive illness such as Seasonal Affective Disorder (SAD). The homogeneous 6-item scale (items 1, 2, 7, 8, 10, and 13) extracted from the HAM-D$_{17}$ (HAMD) [5], the HAM-D$_6$ (HAM$_6$), and its correspondent definite and semidefinite self-rating scales DHAM$_6$, and SHAM$_6$ were also analyzed.

This study also investigated the validity of the self-reported versions of the HAMD, the HAM$_6$, and the MES compared to their corresponding clinician-rated versions in patients with SAD, as well as the sensitivity of these scales to treatment and change. We also examined age-related change in self-rating-ability, patient preference with regard to the different types of questionnaires, score pattern for high scores, and observer-rater reliability.

Methods

The severity of SAD is usually measured by The Structured Guide for the Hamilton Depression Seasonal Affective Disorder (SIGH-SAD) [37]. This scale is a combination of the American HAMD (actually an older version:

Fig. 1 The definite (DHQ/DMQ/DHAM$_6$) and semidefinite (SHQ/SMQ/SHAM$_6$) design for the item “Social activities and interests”

Basic design for the Definite and Semidefinite scales

Definite

DHQ/DMQ/DHAM$_6$

□ a. My daily activities have been as usual.

□ b. I have been a little less interested in my usual activities.

□ c. I have had difficulties performing my usual activities.

□ d. I have had difficulties performing even simple routine activities.

□ e. I have been unable to perform even the most simple activities without help.

Semidefinite

SHQ/SMQ/SHAM$_6$

Has your ability to perform your daily activities been reduced?

□ a. Not at all

□ b. A little

□ c. Clearly

□ d. Very

□ e. Extremely

HAM-D_{21}) and a supplementary 8-item scale for atypical depression symptoms (ADS). Here, we have concentrated on the customary HAMD, the HAM_6, and the MES.

Observer scales

The 1986 version of the 17-item HAMD was used [7]. The MES consists of 11 items, which are operationally defined on a five-point scale. Five of these items are shared with the HAMD (1, 2, 3, 7, and 10) [7].

Self-rating scales

For the present study, the self-reported scales were revised to be clearly concordant with the item-definitions and to avoid any incomprehensible or otherwise problematic semantic phrase. The self-rating scales were thus enhanced but not basically changed. After completing the questionnaires, the patients were asked to state which kind of questionnaire covered their condition best, which was the easier to fill in, and whether they preferred to be interviewed about their condition or would rather fill in a questionnaire.

Patients

A total of 85 patients meeting the DSM-IIIR criteria for SAD were included in our study. The patients were *randomly* recruited from a larger sample of patients with SAD participating in the Martiny et al. study [25] investigating relapse prevention. We only included 85 patients, because we were investigating scale treatment differences and did not want the study power to be too high or to low, as we hoped to find significant placebo/treatment differences for some scales and not for other scales. The patient sample was highly censored at baseline. Only patients scoring a minimum of 13 and maximum of 22 on the HAMD were included. Mean age was 46.4 (SD: 13.3, range 21–75 years), male/female ratio: 22/63. In order not to bias conclusions, all patients in our study had to have paired observer and self-ratings. Two patients had inadequate self-ratings at two time points, and one at four time points, thus the observer ratings for these few patients' time points were omitted. Taking into consideration the large amount of self-ratings without inadequacies and the low amount of instruction patients had received, we found this loss to be very small. No attempt was made to replace missing values. Over time, the patient numbers decreased from 44 to 32 for those receiving placebo (73% completed) and from 41 to 32 for patients receiving citalopram (78% completed). Drop out was not informative; that is we found no kind of systematically missing data. Patients were given no aid other than the questionnaire form instructions: "Would you please fill in this questionnaire to the best of your ability with regard to your condition for the *past 3 days.*" Written informed consent was obtained from all subjects, and the study was approved by the Ethics Committee of Frederiksberg.

Rating procedure

Each patient was rated by expert raters at baseline (Week 0), and at Weeks 1, 3, 9, and 15. The baseline interviews in the main study comprised the HAMD and the MES. To avoid a priming effect of the observer ratings on patients' answers to the questionnaires, the BDI [9] was inserted at all time points between the HAMD/MES interviews and the completion of the patients' definite and semidefinite self-rating scales. The patients were randomly assigned to complete the definite and the semidefinite scales in alternating order, and the items in the SHQ/SMQ were presented in an order opposite to that of the DHQ/DMQ.

Treatment

The patients were treated with morning exposure to bright, full-spectrum light (5,000 Lux) for 2 h each day during the first week. At end of Week 1, patients were randomly (double-blinded) assigned to either 20 mg of citalopram or placebo. The drug treatment could be adjusted in accordance with response to treatment [25].

Statistical analysis

Parametric statistics were used for interval level data, and non-parametric statistics were used for ordinal level data. All scale sum scores were not significantly different from a normal distribution (Kolmogorov–Smirnov test). As a measure of association, correlations are not independent on the score distribution. Even if the difference between two scales' sum scores remains unchanged, the correlation may increase notably with broadening of the variance/range of the sum scores [35]. The semidefinite assessments generally featured ranges which were more than twice as large as those of the observer scales. In addition, the observer ratings were highly censored at baseline, but not at Week 1. Thus, an increase in correlations at Week 1 between observer ratings and self-ratings could be due to an increase in variance/range or to a closer association between these ratings. To test whether the association had changed, we calculated the differences as well as the absolute differences for observer/self-rated scales at baseline and at Week 1.

Mean item scores were compared using the Wilcoxon Signed Rank Test, and these comparisons were performed at baseline, profiling the patient's scores before treatment.

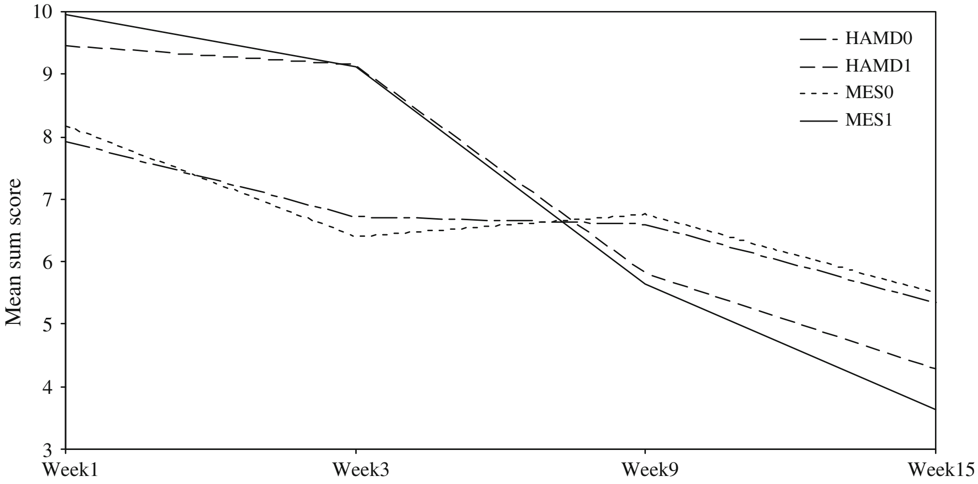

Fig. 2 A means line plot of the HAMD and MES scores split by treatment. HAMD0 = Placebo, HAMD1 = Citalopram, MES0 = Placebo, MES1 = Citalopram

The level of statistical significance was $P = 0.05$. However, when a pair-wise comparison is tested repeatedly over time, it constitutes a test-family (here 5 tests, time point 0, 1, 3, 9, 15) or hypothesis-family, and the risk of type I error becomes inflated. To counteract this *family-wise* (type 1) error rate (FWER), the level of statistical significance or the P value of individual tests were adjusted by the Holm correction, as advocated by Ludbrook [23]. (The Holm correction for multiple tests briefly: The P values are ranged by value. For k tests, the smallest P value is multiplied by k, the next (smallest) P value is multiplied by $k - 1$, and so on.)

The construct validity of the various scales was assessed by the Rasch item-response theory model [2, 21, 32]. We have briefly defined the special Rasch terminology in "Appendix." Item homogeneity and item bias were assessed both locally and globally. A generalized Rasch model was applied, using the item categories present. As our sample was relatively small, two programs were used for the Rasch analysis, namely Digram [21] and Rumm2020 [2]. The two programs supplement each other, as local dependency display is much more elaborated in the Digram, whereas the Rumm2020 has a comprehensive graphical display of item bias. To avoid test inconsistencies, the correction for random significance was revised for the two programs. Digram uses the Benjamini and Hochberg [10] "False Discovery Rate" as default, and this correction was also applied for the Rumm2020 tests. This correction of the very large number of tests made for estimating the acceptance/rejection by the Rasch analysis is recommended [20]. The item-response analyses were performed at Week 1 as advocated by Andrich [2] as the patients sum score ranges at this time point were about twice as large as the baseline scores, *thus providing a much higher power to test the fit of items.*

For the evaluation of scale sensitivity to treatment, the Repeated Mixed Linear Model (RMLM) was used [34]. Inference was provided by the default LS-means test for the time/treatment interaction. The three independent tests at Weeks 3, 9, and 15 were corrected (FWER) ad modum Holm within each scale analysis**.** It is often not recognized that the random assignment of patients to either active treatment or placebo may give one of the treatments an "advantage" if their starting group means are different. The placebo-group mean was considerably lower at Week 1 than the treatment-group mean on all scales, which "favored" the placebo group. We applied the Week 1 scores to the RMLM as a covariate [30]. We were advised to do so and for a relatively strong correlation between Week 1 and outcome variables ($r > 0.35$ on average between Week 1 and Week 15 scores) and a considerable baseline imbalance, it is highly recommended [31] (Fig. 2).

Differences of variance among the observer ratings and the corresponding definite and semidefinite patient ratings were assessed by a method for paired data [3]. The interrater reliability was estimated by The Intraclass Correlation Coefficient (ICC) with an absolute agreement definition [38].

Results

Sum score means

The sum score mean values of the HAMD, HAM_6, and the MES for our subgroup of 85 patients were consistent with the comparable sum scores for the whole group [25].

Table 1 shows the mean sum scores of the various scales at the different time points. Paired comparisons of mean sum scores of the HAMD with the DHQ, the MES with the

Table 1 Mean sum scores and standard deviations (SD) of all scales during the trial period

	Week 0	Week 1	Week 3	Week 9	Week 15
Scales					
HAM_6	11.36 (1.31)	5.31 (3.09)	4.77 (3.86)	3.89 (3.52)	2.80 (3.52)
$DHAM_6$	11.29 (1.70)	5.26 (2.87)	4.68 (3.64)	3.63 (3.33)	2.72 (3.37)
$SHAM_6$	12.56 (2.74)	6.85 (4.16)	5.71 (4.38)	4.63 (4.19)	3.53 (3.81)
HAMD	18.07 (2.52)	8.65 (4.98)	7.88 (5.92)	6.25 (5.19)	4.81 (5.38)
DHQ	18.13 (2.91)	8.86 (5.07)	8.20 (5.95)	6.32 (5.33)	4.72 (5.25)
SHQ	21.51 (5.45)	12.13 (6.78)	10.83 (7.44)	8.91 (6.60)	7.14 (6.40)
MES	18.49 (2.57)	9.02 (5.20)	7.70 (5.70)	6.26 (5.48)	4.56 (5.71)
DMQ	18.55 (3.08)	8.81 (5.17)	7.68 (5.84)	5.83 (5.57)	4.73 (6.00)
SMQ	21.68 (5.60)	11.06 (6.78)	9.18 (7.27)	7.46 (6.82)	5.94 (6.66)
P values for comparisons of mean sum scores and (|) *sum scores variances*					
Compared scales					
$HAM6/DHAM_6$	0.589 | 0.005	0.645 | 0.143	0.604 | 0.171	0.298 | 0.178	0.460 | 0.167
$HAM6/SHAM_6$	0.000 | 0.000	0.000 | 0.000	0.000 | 0.039	0.001 | 0.002	0.000 | 0.037
$DHAM6/SHAM_6$	0.000 | 0.000	0.000 | 0.000	0.000 | 0.000	0.000 | 0.000	0.000 | 0.007
HAMD/DHQ	0.758 | 0.153	0.246 | 0.594	0.141 | 0.922	0.771 | 0.512	0.527 | 0.373
HAMD/SHQ	0.000 | 0.000	0.000 | 0.000	0.000 | 0.000	0.000 | 0.000	0.000 | 0.001
DHQ/SHQ	0.000 | 0.000	0.000 | 0.000	0.000 | 0.000	0.000 | 0.000	0.000 | 0.000
MES/DMQ	0.786 | 0.059	0.229 | 0.067	0.964 | 0.610	0,141 | 0.638	0.287 | 0.192
MES/SMQ	0.000 | 0.000	0.000 | 0.000	0.001 | 0.000	0.000 | 0.000	0.000 | 0.000
DMQ/SMQ	0.000 | 0.000	0.000 | 0.000	0.000 | 0.000	0.000 | 0.000	0.000 | 0.014

DMQ, and the HAM_6 with the $DHAM_6$ showed no significant differences at any time point. The corresponding SHQ, SMQ, $SHAM_6$ sum scores consistently over-scored and were statistically highly significantly different compared to the sum scores of the HAMD, MES, HAM_6, and the DHQ, DMQ, $DHAM_6$.

Sum score variances

The MES and the DMQ sum score variances were not significantly different at any time point, whereas the SMQ sum score variance significantly exceeded both the MES and DMQ variances at all time points (Fig. 3). This pattern was essentially the same for the differences of variances of the HAMD/DHQ/SHQ and $HAM_6/DHAM_6/SHAM_6$. The only exceptions (out of 9 × 5 tests) to this pattern were the $HAM_6/DHAM_6$ variances, which were significantly different at baseline. This exception may be due to the censoring of the HAMD sum score at baseline.

Convergent validity

Table 2 shows that from Week 1, the correlation between all scales was notably higher than at baseline. In particular, the correlations between the observer scales and their corresponding definite scales were very high ($r < 0.9$). The results of paired testing of differences between baseline and Week 1 ratings were non-significant, indicating that the association of the corresponding observer-rating/self-ratings *had not changed* between baseline and Week 1. Even the association of the semidefinite/definite ratings had not changed. The increase in correlations from baseline to Week 1 ratings is thus more likely to be a statistical consequence of the higher ranges and variances at Week 1 than an expression of closer association of the scales. The scale correlations of Week 1 (and Weeks 3, 9, and 15) are probably more reliable than the baseline correlations.

At baseline, the sum score correlations between the different observer scales were much lower than the corresponding correlations of the self-rated definite scales with their parent observer scales. The results of Week 1 indicate that the censoring of HAMD at baseline is the main reason for this.

Comparisons at item level

Figure 4 shows and compares the mean scores of individual items at baseline for the corresponding observer items (HAMD/MES), the definite items (DHQ-DMQ), and the semidefinite items (SHQ-SMQ). In accordance with the semidefinite sum score, the semidefinite items for "mood" ($P = 0.001$), "suicidal impulses" ($P > 0.0001$), "social activities and interests" ($P = 0.0007$), "retardation" ($P = 0.0001$), "agitation" ($P > 0.0001$), "hypochondria"

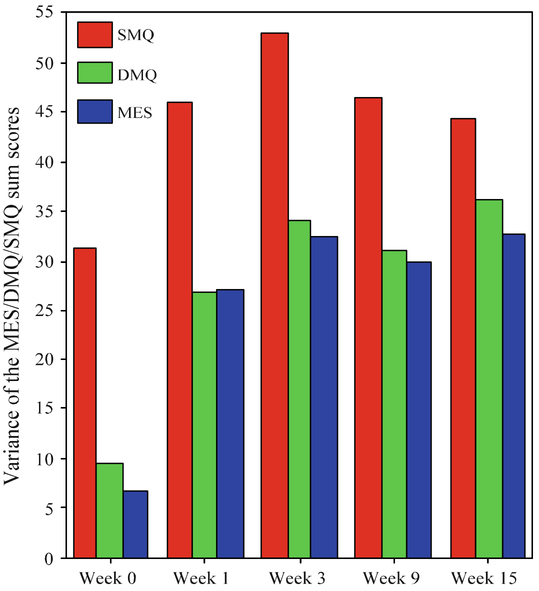

Fig. 3 Basic pattern of comparative scale sum score variances

($P = 0.00001$), "insight" ($P > 0.0001$), "insomnia" ($P = 0.005$) "motor retardation"($P > 0.0001$), and "verbal retardation"($P > 0.0001$) all over-scored significantly in comparison with the corresponding HAMD/MES items.

In contrast, only "suicidal impulses," of the DHQ-DMQ items, scored significantly different (higher, $P = 0.002$) than the corresponding HAMD-MES items. The so-called "observable behavior" items of "agitation," "retardation," and "verbal retardation" scored significantly higher on the semidefinite questionnaires, while there was no significant difference for these items between the corresponding observer and definite items. Most of the semidefinite items had higher mean values than the corresponding mean values of the HAMD/MES. Even if the significantly higher items of the semidefinite scales were deleted from the scale (i.e., HAMD-MES 23 Item scale), the mean sum score of the remaining semidefinite scale still over-scored significantly in comparison with the mean sum scores of the remaining observer scale ($P = 0.017$) and the remaining definite scale ($P = 0.0004$), i.e., *over-scoring on the semidefinite scales was a general phenomenon*. The HAMD items "hypochondriasis" and "insight" had a baseline mean of 0.02 and 0, respectively. The corresponding definite item scores were not different from the observer items. The significant over-scoring of the semidefinite corresponding items on these *redundant* items should be noted (Fig. 4).

Construct validity and reliability

The item responses of the MES, DMQ, HAM_6, and $DHAM_6$ were all accepted by the Rasch analysis. The items fitted the model and were without local dependencies. Items were homogenous for high- and low-score groups and unbiased in relation to age and gender both locally and globally *that is, the sum score of these scales exhibited unidimensionality*. The Rumm2020 analysis stated that the power for testing item fit analysis was *"excellent."* These scales will be referred to as the valid scales.

The HAMD, DHQ, SHQ, SMQ, and $SHAM_6$ item responses were all rejected by the Rasch analysis. On the HAMD, the redundant item "insight" was expelled by both programs and the analysis continued without this item. The HAMD was rejected globally, and the P value was exactly the same for both programs, namely $P = 0.002$. A local dependency between "psychic anxiety" and "somatic anxiety" was prominent, and so significant ($P < 0.0000$) that the two items rather functioned as one item (a so-called item bundle), thus rendering no extra information for the sum score. Exactly the same highly significant

Table 2 Correlations among the various scales

	Week 0	Week 1	Week 3	Week 9	Week 15
HAMD/MES	0.63	0.95	0.93	0.94	0.96
HAMD/HAM_6	0.57	0.94	0.95	0.95	0.96
MES/HAM_6	0.74	0.95	0.94	0.95	0.97
Correlations among corresponding observer scales and questionnaires					
HAM_6/$DHAM_6$	0.71	0.95	0.92	0.94	0.97
HAM_6/$SHAM_6$	0.45	0.85	0.88	0.90	0.95
$DHAM_6$/$SHAM_6$	0.50	0.82	0.95	0.93	0.93
HAMD/DHQ	0.80	0.95	0.95	0.93	0.98
HAMD/SHQ	0.65	0.85	0.89	0.87	0.92
DHQ/SHQ	0.64	0.85	0.94	0.91	0.92
MES/DMQ	0.77	0.95	0.91	0.95	0.98
MES/SMQ	0.60	0.84	0.85	0.92	0.95
DMQ/SMQ	0.57	0.87	0.95	0.95	0.94

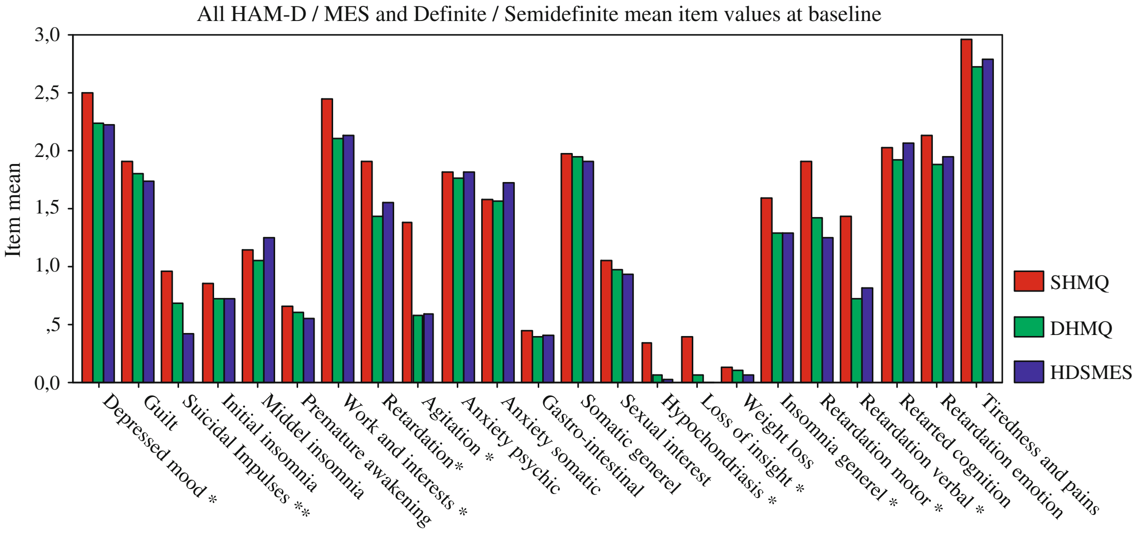

Fig. 4 Comparisons of the HAMD-MES, the Definite and the Semidefinite items at baseline. *Semidefinite item means are significantly greater than corresponding HAMD-MES. **Semidefinite and Definite item means are significantly greater than corresponding HAMD-MES item means. The first 17 items constitute the HAMD. Items 1, 2, 3, 7, 10, 18, 19, 20, 21, 22, 23, constitute the MES

($P < 0.0000$) *redundancy* was found with the corresponding two DHQ items.

The item "guilt" of the semidefinite scales was significantly biased. Men scored locally and globally higher than women on this item. SHQ had the poorest global fit of all scales ($P > 0.0001$) as estimated by both programs and also had a large number of local dependencies. Surprisingly, the item "mood" fitted the model poorly. The SMQ was rejected globally ($P = 0.02$), the "insomnia" item fitted the model poorly, and the SMQ exhibited a large number of local dependencies, including "guilt" and "social activities and interests," and the item "verbal retardation" was locally dependent with "motor retardation" and "sleep." The SHAM$_6$ was rejected because of the local bias in relation to "guilt" ($P = 0.002$), and a severe local dependence of "guilt" and "social activities and interests." It was also globally biased with gender. Generally, there were no consistent problems with the items "appetite" and "weight loss," as might have been expected for patients with SAD. The Digram estimated the test true-score correlation to be above 0.9 for the valid scales, indicating a highly satisfactory *reliability* for these scales.

The two programs produced only minor differences and demonstrated remarkably consistent results.

Sensitivity to treatment

Sum scores for actively treated patients and patients receiving placebo were compared over time by RMLM. The Holm (FWER) corrected P values (3 tests) for the LS-means differences of these group sum scores of each scale at Week 15 were: HAMD: $P = 0.09$, HDQ: $P = 0.17$, SHQ: $P = 0.62$, *MES:* $P = 0.012$, *DMQ:* $P = 0.03$, SMQ: $P = 0.18$, *HAM$_6$:* $P = 0.02$, *DHAM$_6$:* $P = 0.03$, SHAM$_6$: $P = 0.24$. A significant drug–placebo difference was thus identified by the MES and HAM$_6$ observer ratings and the corresponding definite self-ratings (DMQ and DHAM$_6$), but not with the HAMD or any of the semidefinite scales. No significant drug–placebo was found at time point 9 for any scale.

Sensitivity to change

The corresponding mean values and variances of the HAM$_6$/DHAM$_6$, HAMD/DHQ, and MES/DMQ showed non-significant differences between the observer scale and the corresponding definite self-rating scale at all time points, while the corresponding parameters of the SHAM$_6$/SHQ/SMQ were significantly higher (Table 1). The repeated testing of changes in patient's depression level, with steadily decreasing mean sum scores and increasing variances thus showed similar sensitivity to change for the definite self-rating-scales DHAM$_6$/DHQ/DMQ as found for their parent observer scales.

Patients with high sum scores

For a subgroup of patients scoring 19–25 on the MES at baseline ($N = 40$), the mean sum score and variances of corresponding observer and definite scale sum scores were (as for the total groups) not significantly different. The sum score means and variances of the SMQ, SHQ, and SHAM$_6$ highly significantly exceeded the sum score means and variances of *both* the corresponding observer scales and the corresponding definite scales (SMQ/MES $P = 0.000$, SMQ/DMQ $P = 0.000$, MES/DMQ $P = 0.548$, SHQ/HAMD $P = 0.000$, SHQ/DHQ $P = 0.000$, HAMD/DHQ $P = 0.936$, SHAM$_6$/HAM$_6$ $P = 0.000$, SHAM$_6$/DHAM$_6$ $P = 0.001$, HAM$_6$/DHAM$_6$ $P = 0.212$).

Age and self-rating ability

Age had a high range (21–75 years mean: 46.34 SD: 13.29). We used simple regression of age with the differences between the observer scales and the corresponding self-reported scales. There was no significant association between these differences and age for any scale. This result indicates that the self-rating ability did not decline with age for these patients.

Preference

According to self-rated patient opinion of the questionnaires at baseline, the definite questionnaires covered the patients' conditions best (76% in favor of DHQ/DMQ; 24% in favor of the SHQ/SMQ, $P < 0.001$). The semidefinite questionnaires were found to be easier to complete by 64% of the patients, while 36% found the definite questionnaires easier to complete, but the difference was not statistically significant (NS). A relatively large group preferred to self-rate their symptoms (44%), while 56% of patients preferred an interview (NS).

Inter-rater reliability

Patients were rated by 3 expert raters, each with more than 15 years of experience with these scales. The ICC was based on 11 joint ratings, with a HAMD of 23, 3 (SD 4.4). The inter-rater reliability was 0.95 for the HAMD, 0.94 for the HAM$_6$, and 0.94 for the MES. There were no systematic differences among the raters.

Discussion

The present results of the Rasch analysis for the MES and HAMD are in agreement with an older study by Maier et al. [24] who found that the HAMD would improve if the items "hypochondriasis" and "insight" were not included. Maier et al. [24] also found the MES item responses to be homogeneous and unbiased using the Rasch model, whereas the HAMD was rejected for lack of item homogeneity. Recently, Bagby et al. [4] convincingly confirmed the inadequacy of the HAMD and advocated the retirement of the HAMD. The high psychometric ability of the HAM$_6$ is supported by the results of other studies [22, 28].

The acceptance by the Rasch analysis of the DMQ and DHAM$_6$ item responses was unequivocal. The sum scores of the DMQ and DHAM$_6$ were thus found to exhibit unidimensional properties. To our knowledge, this has not previously been demonstrated for self-rating scales for depression.

The DHAM$_6$ and SHAM$_6$ were not investigated in Bent-Hansen et al. [11] which means that *this study is the first* to perform item-response theory model validation ad modum Rasch [32] of these scales, to compare their sum scores with the sum scores of the HAM$_6$ and to investigate their sensitivity to treatment and other factors.

Our data show that the correlations between the sum scores of the observer scales and the corresponding definite self-rating scales at Week 1 were more valid than the corresponding correlations at baseline, as the differences between patient and observer-rater did not change between baseline and Week 1, whereas the sum score ranges and variances did. These results are in accordance with the findings of Senra and Polaino [35], who concluded that the statistical significant improvements after basement and later on in total score correlations are not due to improved concordance between clinician and patient but are a statistical *artifact* due to the increase in total score ranges.

At Week 1 (and at Weeks 3, 9, and 15), the correlations between the definite self-rating and observer scales were close to absolute agreement, indicating an almost perfect convergent validity. Our results have confirmed that low anchoring of item-response options (semidefinite format) is of major importance in explaining self-report over-scoring, which many authors have seen as an unavoidable drawback of self-rating [12, 16]. We have also shown that the sum scores of the semidefinite scales exhibit large error variances compared to the sum score variances of the definite scales. The whole idea of constructing the definite scales could be described as an effort to restrict the variance of a self-rating scale and to limit over-scoring. The method was obviously successful as we found essentially no differences as regards sum score means and variances between definite self-rating scales and corresponding observer scales.

The link between the practical assessment and the theoretical analysis was established, since a high sensitivity to treatment was shown only for the valid scales. High scale sensitivity to treatment has the important implication that fewer patients need to participate in controlled trials.

Increasing the level of anchoring of observer scales may also increase their reliability. When an observer-rated scale is standardized, ensuring a better anchoring of question options, the reliability increases as a result of lower inter-rater variance [26]. Manuals for observer-rated scales can also be regarded as anchoring instruments, their purpose being to reduce the variance of the observer ratings. We demonstrated a very high inter-rater reliability in the present study. This is obviously of great importance to avoid increasing and/or biasing the observer scale sum score variance. Nevertheless, information on inter-rater reliability of clinical trials is often lacking [27].

With regard to individual items, the only differences between the MES/HAMD and the DHQ/DMQ were noted with respect to the items "suicidal impulses." Raters may have under-scored the item "suicidal impulses" as patients might be reluctant to reveal the extent of these thoughts to an interviewer but feel freer when filling in a questionnaire. It is not clear if this is true or not, but the problem would have been unnoticed if only observer ratings had been used. It is a good example of the supplementary information provided by the inclusion of patients self-rating. The use of multilevel information in clinical trials is gaining general recommendation [29].

We have often noticed that patients are keenly aware of physical retardation, agitation, and particularly verbal retardation. The data analysis shows that patient ratings on the definite scales as regards physical, verbal, emotional, and cognitive retardation did not differ significantly from observer ratings. If a depressed patient expresses an uncommon restlessness (agitation), which is not observable, the experienced rater is supposed to rate zero on this item, but this places the rater in a difficult position and seems to entail unnecessary loss of information.

Importantly, we noted no decline in self-rating ability with age. A significant proportion of patients found that the definite items covered their condition more adequately than the items of the semidefinite scales.

Our results are limited to outpatients with moderate depression. However, the high-scoring sub-group in the present study was also validly assessed by the MES, DMQ, HAM_6, and $DHAM_6$ scales, but these patients are only at the high end of "moderately depressed."

Conclusion

We have shown that semidefinite answering options in self-rating questionnaires are not as powerful from a psychometrical point of view as anchored (definite) options.

In all comparisons, the sum scores of definite self-rating scales (DHQ, DMQ, $DHAM_6$) did not differ from the sum scores of their parent observer scales HAMD, MES, and HAM_6. Furthermore, the variances of the paired observer ratings and definite scores were essentially not different from each other, and the convergent validity was close to identical. To the best of our knowledge, this close correspondence between observer ratings and patient ratings has not been demonstrated before. However, the results are consistent with our earlier findings [11] and the results of this more comprehensive investigation, with 5 time points as opposed to one, confirm that the sum scores of the definite and semidefinite self-reported scales generally provide the same results, even in SAD, a different category of major depression.

Among the self-reported scales with high convergent validity, only the DMQ and the $DHAM_6$ could be confirmed by the Rasch analysis to have construct validity together with high sensitivity to treatment along with their corresponding observer scales (the MES and the HAM_6). These self-report scales (and their parent observer scales) were able to identify significant treatment differences that were not detected by the non-valid scales. Notable among the latter was the observer-rated HAMD, for years the gold standard scale.

We have identified two very interesting new self-rating instruments: The DMQ and the $DHAM_6$—both with very high psychometric properties. To investigate their performance further, we plan to compare these scales with the BDI—the most frequently used self-rating for assessing levels of depression.

Acknowledgments We thank the Statistical Service of the Southern Danish University for providing professional guidance with revising and supervising statistical procedures and interpretations of results in this manuscript. The authors thank Professor Svend Kreiner for giving informative seminars in scale validation—with emphasis on understanding the Rasch model, its practical application, and interpretation using Digram.

Conflict of interest statement Bent-Hansen J: None. Bech P: Over the past 3 years until August 2008, Per Bech has occasionally received funding from and been speaker or member of advisory boards for pharmaceutical companies with an interest in the drug treatment of affective disorders (Astra-Zeneca, Lilly, H. Lundbeck A/S, Organon).

Appendix

A brief definition of important terms of the Rasch model (as defined for DIGRAM users)

The Rasch model is the simplest model within item-response (IRT) models. It assumes that there is a monotonic, increasing or decreasing consistency between the patient's condition and symptoms prevalence.

The dramatic data reduction represented by the sum score of a scale should be assessed for sum score validity. Item responses are assumed to have a high inter-correlation across the whole patient-population, but for a given sum score group (conditionally), item responses should be independent (*local independence*) and without *bias* with regard to different patient groups (in this study, gender and different age groups). Item responses should not differ significantly from expected scores for high-scoring and low-scoring patients (*global homogeneity*), and they should not differ significantly from expected scores for men and women, low and high groups of age (*global bias*).

By analyzing item responses locally and globally, the Rasch analysis tests whether the sum score of a scale is validly measuring, what it is supposed to measure. If that is so, the scale is only measuring one dimension—it has unidimensional properties.

References

1. Ajdacic-Gross V, Vetter S, Müller M, Kawohl W, Frey F, Lupi G, Blechschmidt A, Born C, Latal B, Rössler W (2009) Risk factors for stuttering: a secondary analysis of a large data base. Eur Arch Psychiatry Clin Neurosci Published online: 15 Oct 2009
2. Andrich D (2004) The interpreting RUMM2020 monograph. Part I: PDF file-downloadable from http://www.rummlab.com.au/. Paragraph 3.7. Chronbach's Alfa and traditional test theory
3. Armitage P, Berry G, Matthews JNS (2002) Statistical methods in medical research, 4th edn. Blackwell Scientific Publications, London, pp 203–204
4. Bagby RM, 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
5. Bech P (1981) Rating scales for affective disorders: their validity and consistency. Thesis. Acta Psychiatr Scand 64(Suppl 295):1–101
6. Bech P, Kastrup M, Rafaelsen OJ (1986) Mini-compendium of rating scales for states of anxiety, depression, mania and schizophrenia with corresponding DSM-III syndromes. Acta Psychiatr Scand 73(Suppl 326):1–37
7. Bech P (1993) Rating scales for psychopathology, health status and quality of life: a compendium on documentation in accordance with the DSM 3-R and WHO systems. Springer, Berlin
8. Bech P (2002) The Bech–Rafaelsen melancholia Scale (MES) in clinical trials of therapies in depressive disorders: a 20-year review of its use as outcome measure. Acta Psychiatr Scand 106:252–264
9. Beck AT, Ward CH, Mendelson M, Mock ErbaughJ (1961) An inventory for measuring depression. Arch Gen Psychiatry 4:561–571
10. BenjaminiY HochbergY (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B 57:289–300
11. Bent-Hansen J, Kørner A, Lauritzen L, Clemmesen L, Lunde M (1995) A definite and a semidefinite questionnaire version of the Hamilton/Melancholia (HAMD/MES) scale. J Affect Disord 33:143–150
12. Caroll BJ, Feinberg M, Smouse PE, Rawson SG, Greden JF (1981) The Carroll Rating Scale for depression 1. Development, reliability and validation. Br J Psychiatry 138:194–200
13. Demyttenaere K, Fruyt JD (2003) Getting what you ask for: on the Selectivity of Depression Rating Scales. Psychother Psychosom 72:61–70
14. Domken M, Scott J, Kelly P (1994) What factors predict discrepancies between self- and observer-ratings of depression? J Affect Disord 31:253–259
15. Faravelli C, Albanesi G, Poli E (1986) Assessment of depression: a comparison of rating scales. J Affect Disord 11:245–253
16. Feinberg M, Carroll BJ, Smouse P, Rawson S (1981) The Caroll Rating Scale for depression 3. Comparison with other rating instruments. Br J Psychiatry 138:205–209
17. Hamilton M (1960) A rating scale for depression. J Neurol Neurosurg Psychiatry 23:56–62
18. Hamilton M (1967) Development of a rating sale for primary depressive illness. Br J Soc Clin Psychol 6:278–296
19. Kellner R (1986) The symptom rating test. In: Sartorius N, Ban TA (eds) Assessment of depression. Springer, Berlin, pp 213–220
20. Keselman HJ, Cribbie R, Holland B (2002) Controlling the rate of Type I error over a large set of statistical tests. Br J Math Stat Psychiatry 55:27–39
21. Kreiner S (2003) Introduction to DIGRAM. Biostatistical Dept., University of Copenhagen http://www.pubhealth.ku.dk/upload/application/msword/f51d6748/rr-03-10.doc
22. Lecrubier Y, Bech P (2007) The Ham D_6 is more homogeneous and as sensitive as the $HamD_{17}$. European Psychiatry 22:252–255
23. Ludbrook J (1998) Multiple comparison procedures updated. Clin Exp Pharmacol Physiol 25:1032–1037
24. Maier W, Philipp M, Heuser I (1988) Improving depression severity assessment-I. Reliability internal validity and sensitivity to change of three observer depression scales. J Psychiatr Res 22:3–12
25. Martiny K, Lunde M, Simonsen C, Clemmesen L, Poulsen DL, Solstad K, Bech (2004) Relapse prevention by citalopram in SAD patients responding to 1 week of light therapy. A placebo-controlled study. Acta Psychiatr Scand 109:230–234
26. Miller PR, Dasher R, Collins R, Griffiths P, Brown F (2001) Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. Psychiatry Res 105:255–265
27. Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG (2002) Interrater reliability in clinical trials of depressive disorders. Am J Psychiatry 159:1598–1600
28. Möller HJ (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
29. Möller HJ, Broich K (2009) Principle standards and problems regarding proof of efficacy in clinical psychopharmacology. Eur Arch Psychiatry Clin Neurosci Published online: 04 Nov 2009
30. Overall JE, Doyle SR (1994) Implications of chance baseline differences in repeated measurement designs. J Biopharm Stat 4:199–216
31. Pocock SJ, Assmann SE, Enos LE, Kasten LE (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 21:2917–2930
32. Rasch G (1980) Probabilistic models for some intelligence and attainment tests. University of Chicago Press, Chicago
33. Rush AJ, Hiser W, Giles D (1987) A comparison of self-reported versus clinician-rated symptoms in depression. J Clin Psychiatry 48:246–248
34. SAS Documentation (software version 9.1). The mixed linear model for repeated data
35. Senra C, Polaino A (1993) Concordance between clinical and self-report depression scales during the acute phase and after treatment. J Affect Disord 27:13–20
36. Tondo L, Burrai C, Scamonatti L, Weissenburger J, Rush AJ (1988) Comparison between clinician-rated and self-reported depressive symptoms in Italian psychiatric patients. Neuropsychobiology 19:1–5
37. Williams JB, Link MJ, Rosenthal NE, Terman M (1988) Structured Interview guide for the Hamilton Depression rating Scale, Seasonal Affective Disorders Version (SIGH-SAD). New York Psychiatric Institute, New York
38. Winer BJ (1971) Statistical principles in experimental design, 2nd edn. McGrawHill, New York, pp 289–296