**Neuropsychopharmacology (2005) 30,** 405–416

**www.neuropsychopharmacology.org**

# Self-Reported Depressive Symptom Measures: Sensitivity to Detecting Change in a Randomized, Controlled Trial of Chronically Depressed, Nonpsychotic Outpatients

**A John Rush*[,1], Madhukar H Trivedi[1], Thomas J Carmody[2], Hisham M Ibrahim[1], John C Markowitz[3,4], Gabor I Keitner[5], Susan G Kornstein[6], Bruce Arnow[7], Daniel N Klein[8], Rachel Manber[7], David L Dunner[9], Alan J Gelenberg[10], James H Kocsis[3], Charles B Nemeroff[11], Jan Fawcett[12], Michael E Thase[13], James M Russell[14], Darlene N Jody[15], Frances E Borian[15] and Martin B Keller[16]**

*[1]Department of Psychiatry, University of Texas Southwestern Medical Center, Dallas, TX, USA; [2]Academic Computing Services, University of Texas Southwestern Medical Center, Dallas, TX, USA; [3]Department of Psychiatry, Cornell University Medical College, New York, NY, USA; [4]Department of Psychiatry New York State Psychiatric Institute, New York, NY, USA; [5]Department of Psychiatry, Brown University and Rhode Island Hospital, Providence, RI, USA; [6]Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA; [7]Department of Psychiatry & Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA; [8]Department of Psychology, State University of New York at Stony Brook, USA; [9]Department of Psychiatry & Behavioral Sciences, University of Washington, Seattle, WA, USA; [10]Department of Psychiatry, University of Arizona Health Sciences Center, Tucson, AZ, USA; [11]Department of Psychiatry & Behavioral Sciences, Emory University School of Medicine, Atlanta, GA, USA; [12]Department of Psychiatry, Rush-St Luke's-Presbyterian Medical Center, Chicago, IL, USA; [13]Department of Psychiatry, Western Psychiatric Institute & Clinic, University of Pittsburgh Medical Center, Pittsburgh, PA, USA; [14]Department of Psychiatry & Behavioral Sciences, University of Texas Medical Branch, Galveston, TX, USA; [15]Bristol-Myers Squibb, Plainsboro, NJ, Atlanta, GA, USA; [16]Department of Psychiatry & Human Behavior, Brown University and Butler Hospital, Providence, RI, USA*

This study evaluated and compared the performance of three self-report measures: (1) 30-item Inventory of Depressive Symptomatology-Self-Report ($IDS\text{-}SR_{30}$); (2) 16-item Quick Inventory of Depressive Symptomatology-Self-Report ($QIDS\text{-}SR_{16}$); and (3) Patient Global Impression-Improvement (PGI-I) in assessing clinical outcomes in depressed patients during a 12-week, acute phase, randomized, controlled trial comparing nefazodone, cognitive-behavioral analysis system of psychotherapy (CBASP), and the combination in the treatment of chronic depression. The $IDS\text{-}SR_{30}$, $QIDS\text{-}SR_{16}$, PGI-I, and the 24-item Hamilton Depression Rating Scale ($HDRS_{24}$) ratings were collected at baseline and at weeks 1–4, 6, 8, 10, and 12. Response was defined *a priori* as a $\geqslant$50% reduction in baseline total score for the $IDS\text{-}SR_{30}$ or for the $QIDS\text{-}SR_{16}$ or as a PGI-I score of 1 or 2 at exit. Overall response rates (LOCF) to nefazodone were 41% ($IDS\text{-}SR_{30}$), 45% ($QIDS\text{-}SR_{16}$), 53% (PCI-I), and 47% ($HDRS_{17}$). For CBASP, response rates were 41% ($IDS\text{-}SR_{30}$), 45% ($QIDS\text{-}SR_{16}$), 48% (PGI-I), and 46% ($HDRS_{17}$). For the combination, response rates were 68% ($IDS\text{-}SR_{30}$ and $QIDS\text{-}SR_{16}$), 73% (PGI-I), and 76% ($HDRS_{17}$). Similarly, remission rates were comparable for nefazodone ($IDS\text{-}SR_{30} = 32\%$, $QIDS\text{-}SR_{16} = 28\%$, PGI-I = 22%, $HDRS_{17} = 30\%$), for CBASP ($IDS\text{-}SR_{30} = 32\%$, $QIDS\text{-}SR_{16} = 30\%$, PGI-I = 21%, $HDRS_{17} = 32\%$), and for the combination ($IDS\text{-}SR_{30} = 52\%$, $QIDS\text{-}SR_{16} = 50\%$, PGI-I = 25%, $HDRS_{17} = 49\%$). Both the $IDS\text{-}SR_{30}$ and $QIDS\text{-}SR_{16}$ closely mirrored and confirmed findings based on the $HDRS_{24}$. These findings raise the possibility that these two self-reports could provide cost- and time-efficient substitutes for clinician ratings in treatment trials of outpatients with nonpsychotic MDD without cognitive impairment. Global patient ratings such as the PGI-I, as opposed to specific item-based ratings, provide less valid findings.
*Neuropsychopharmacology* (2005) **30,** 405–416, advance online publication, 1 December 2004; doi:10.1038/sj.npp.1300614

**Keywords:** chronic depression; psychotherapy; nefazodone; symptom measures

*Correspondence: Dr AJ Rush, Department of Psychiatry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390-9086, USA, Tel: 214 648 4601, Fax: 214 648 4612, E-mail: john.rush@utsouthwestern.edu
Received 17 June 2004; revised 16 September 2004; accepted 28 September 2004
Online publication: 11 October 2004 at http://www.acnp.org/citations/NPP101104040282/default.pdf

## INTRODUCTION

Most randomized, controlled trials (RCTs) of medication or of psychotherapy rely on clinician ratings of symptoms to assess treatment effects and to define both response and remission (Depression Guideline Panel, 1993). The 17- and 21-item versions of the Hamilton Depression Rating Scale (HDRS) (Hamilton, 1960, 1967) and the 10-item

Montgomery-Åsberg Depression Rating Scale (MADRS) (Montgomery and Asberg, 1979) are the most commonly used clinical ratings in RCTs of adults with major depressive disorder (MDD) (Yonkers and Samson, 2000), although neither scale rates all of the nine criterion symptom domains needed to diagnose a major depressive episode. The heavy reliance on clinician ratings as opposed to self-report ratings such as the Beck Depression Inventory (BDI) (Beck *et al*, 1961, 1979), the Zung Self-Rating Scale (ZSRS) for depression (Zung, 1965, 1986), or the Carroll Rating Scale (CRS) for depression (Carroll *et al*, 1981) may be due to several factors, including: (1) unavailability of self-reports during early clinical trials, (2) concerns that self-reports may be biased by patients' expectations of improvement, (3) regulatory agency preferences for using clinician ratings, and (4) in some cases, copyright issues. Thus, if one rating were to be valued above all others, many would argue that clinician ratings are more 'valid' and potentially more sensitive to change than self-reports.

This assumption, however, has been challenged (Greenberg *et al*, 1992). These authors have argued that clinicians involved in placebo-controlled medication RCTs may detect subtle cues, thereby 'knowing' which patients are on active medication (Hughes and Krahn, 1985; Rabkin *et al*, 1986; Stallone *et al*, 1975). They contend that this argument is supported by the fact that in some studies only clinician ratings, but not self-reports, distinguish 'active' medication from placebo (Edwards *et al*, 1984; Lambert *et al*, 1986). This perspective suggests that self-reports might be more valid than clinician ratings.

Self-report measures have other strengths that argue for reconsidering their role in both clinical trials and other settings. From a research perspective, the move toward large, multisite, 'simple' effectiveness trials such as the Texas Medication Algorithm Project (TMAP) (Rush *et al*, 1999a, b, 2003a; Trivedi *et al*, 2004a; Biggs *et al*, 2000) or the NIMH-sponsored Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study (Fava *et al*, 2003; Rush *et al*, 2004) would be facilitated by valid self-reports of depressive symptoms because they would reduce the time, effort, and costs incurred in large-scale clinical trials. Additionally, if self-report instruments were shown to accurately reflect clinical status, clinicians in practice might use them to help patients manage their long-term disease and determine what action(s) might be needed (eg visit the clinician to consider dose or treatment change).

Items on self-report scales, such as the BDI or ZSRS, commonly differ from items on clinician ratings such as the HDRS or MADRS. Correlations between the BDI and HDRS range from 0.61 to 0.86, and for the ZSRS and HDRS, they range from 0.56 to 0.79 (Yonkers and Samson, 2000). Correlations between the HDRS and CRS, which have similar items to the HDRS, are slightly higher (0.80) (Bech, 1992).

In addition to differences in the items themselves, other factors might account for the apparent discrepancies between clinician and patient ratings. It has been suggested that cognitive change items are among the last to improve with otherwise effective medication treatment (Prusoff *et al*, 1972). If true, self-reports of symptom severity, especially those containing a substantial number of 'cognitive items,' may be less sensitive than clinician ratings in detecting symptomatic changes, especially in treatment trials of short duration. A self-report that matches the items on the clinician rating would allow a more accurate test of Greenberg *et al*'s (1992) suggestion.

The 30-item Inventory of Depressive Symptomatology-Self-Report ($IDS\text{-}SR_{30}$) and the matched clinician rating ($IDS\text{-}C_{30}$) (Rush *et al*, 1986, 1996) were developed to carefully assess all core criterion diagnostic depressive symptoms, as well as all Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (American Psychiatric Association, 1994) atypical symptom features (eg weight gain, appetite increase, interpersonal rejection sensitivity, hypersomnia, and leaden paralysis), and DSM-IV melancholic symptom features (eg anhedonia, unreactive mood, distinct quality to mood, etc.). The $IDS\text{-}C_{30}$ and $IDS\text{-}SR_{30}$ have identical items. Both exclude uncommon symptoms (ie obsessions and compulsions, lack of insight, paranoia/suspiciousness, depersonalization/derealization) that are included in the 21-item version of the HDRS ($HDRS_{21}$). Thus, the item content of both the $IDS\text{-}C_{30}$ and $IDS\text{-}SR_{30}$ are consistent with recent recommendations to exclude uncommonly encountered symptoms (Mazure *et al*, 1986; Gibbons *et al*, 1993; Snaith, 1993; Grundy *et al*, 1994; Gullion and Rush, 1998).

Evidence of acceptable psychometric properties of the $IDS\text{-}C_{30}$ and $IDS\text{-}SR_{30}$ in depressed outpatients (Rush *et al*, 1986, 1996, 2000, 2003b; Gullion and Rush, 1998; Trivedi *et al*, 2004b) and depressed inpatients (Corruble *et al*, 1999a, b) has been reported. There is also a substantial correlation between total scores on the $IDS\text{-}C_{30}$, $IDS\text{-}SR_{30}$, and $HDRS_{17}$ (Rush *et al*, 1986, 1996; Gullion and Rush, 1998). IDS ratings have been shown to differentiate endogenous from nonendogenous forms of depression (Rush *et al*, 1987; Domken *et al*, 1994), dysthymic disorder from MDD (Rush *et al*, 1987), depressed from nondepressed radiation oncology patients (Jenkins *et al*, 1998), and depressed from nondepressed cocaine-dependent jail inmates (Surís *et al*, 2001).

There is substantial evidence of the correspondence between individual items and total scores on the $IDS\text{-}C_{30}$ and $IDS\text{-}SR_{30}$ (Rush *et al*, 1986, 1996; Tondo *et al*, 1988; Corruble *et al*, 1999a, b; Biggs *et al*, 2000; Trivedi *et al*, 2004b). Therefore, it is useful to examine whether the $IDS\text{-}SR_{30}$ or a shorter version (16-item Quick Inventory of Depressive Symptomatology-Self-Report) ($QIDS\text{-}SR_{16}$) (Rush *et al*, 2000, 2003b; IsHak *et al*, 2002; Trivedi *et al*, 2004b) might provide an alternative, sensitive means to evaluate comparative outcomes in RCTs and other settings. Recent findings by others (eg Carpenter *et al*, 2002) suggest that in small trials, the $IDS\text{-}SR_{30}$ is sensitive to change and is comparable to the $HDRS_{17}$.

Data from a recently reported, large, multicenter, controlled trial (Keller *et al*, 2000) provided an opportunity to evaluate the performance of several self-reports of overall depressive illness severity. This 12-week acute phase, randomized, controlled trial compared nefazodone, cognitive-behavioral analysis system of psychotherapy (CBASP) (McCullough, 2000), and the combination of both nefazodone and CBASP in the treatment of outpatients with chronic, nonpsychotic MDD. This dataset allowed for this *post hoc* comparison of the performance of clinician-rated $HDRS_{24}$ with the $IDS\text{-}SR_{30}$, $QIDS\text{-}SR_{16}$, and the Patient

Global Impression-Improvement (PGI-I) (adapted from the Clinical Global Impression-Improvement) (CGI-I) (Guy, 1976) in comparing outcomes with three different treatments (nefazodone, CBASP, or combined nefazodone and CBASP).

## MATERIALS AND METHODS

### Subjects

Institutional review boards at each of the 12 participating sites approved the study. Written informed consent was obtained from all 681 outpatient participants. The methods and overall outcomes of this 12-week acute phase, three-cell RCT have been detailed elsewhere (Keller *et al*, 2000). The Structured Clinical Interview for DSM-IV™ Axis I Disorders (SCID-I) (First *et al*, 1997) established the diagnosis. All participants, 18–75 years of age, met DSM-IV criteria for a major depressive episode that was (1) of at least 2 years in duration, (2) superimposed on antecedent dysthymia (double depression), or (3) recurrent with incomplete interepisode recovery and total illness duration of $\geqslant 2$ years. The $HDRS_{24}$ score was $\geqslant 20$ at study entry. (See Keller *et al*, 2000 for inclusion/exclusion criteria.)

### Treatment

Nefazodone (b.i.d.) was titrated from 200 mg/day during the first week to 300 mg/day during week 2, followed by weekly dose adjustments in 100 mg/day increments (up to 600 mg/day) to achieve maximum efficacy and acceptable tolerability. CBASP was implemented according to a treatment manual (McCullough, 2000) with twice-weekly sessions (until week 4) and weekly sessions thereafter until week 12.

### Outcome Measures

The present study is based on the $IDS\text{-}SR_{30}$, the $QIDS\text{-}SR_{16}$ that was derived from the $IDS\text{-}SR_{30}$, and the PGI-I. The PGI-I, $IDS\text{-}SR_{30}$, $QIDS\text{-}SR_{16}$, and 24-item HDRS ($HDRS_{24}$) (Miller *et al*, 1985) were obtained at each treatment visit, with the latter being acquired by an evaluator blind to treatment assignment. There were no specific instructions as to the order of test administration. Thus, the PGI-I, $IDS\text{-}SR_{30}$, and $HDRS_{24}$ could have been given in any order depending on clinical convenience.

The $HDRS_{24}$ was the primary outcome measure in the Keller *et al* (2000) RCT. The $HDRS_{24}$ includes 24 items, each rated on a 0–2, 0–3, or 0–4 scale (range = 0–75). The $HDRS_{24}$ was obtained by trained and certified assessors who were masked to treatment assignment, and who were not asked to review nor were they specifically provided patient self-reports prior to obtaining the $HDRS_{24}$ (though some may have had access to these self-reports). The $HDRS_{24}$, as well as the $IDS\text{-}SR_{30}$ and the PGI-I, were obtained at weeks 0, 1, 2, 3, 4, 6, 8, 10, and 12.

The $IDS\text{-}SR_{30}$ includes 30 items, of which 28 are rated on a 0–3 scale. Both weight gain (or loss) and appetite increase (or decrease) are rated at any occasion, which results in 28 items being rated (range of 0–84). Item total correlations, Cronbach's alpha (Cronbach, 1951), and measures of concurrent validity have established acceptable psychometric properties for the $IDS\text{-}SR_{30}$ (Rush *et al*, 1986, 1996; Gullion and Rush, 1998; Corruble *et al*, 1999a; Trivedi *et al*, 2004b).

The $QIDS\text{-}SR_{16}$ contains 16 items that were selected in this study from items provided by the $IDS\text{-}SR_{30}$ to evaluate the overall severity of each of the nine criterion symptom domains used to diagnose DSM-IV MDD (ie sleep, appetite/weight, concentration, energy, sad mood, suicidal ideation, self-concept/guilt, psychomotor activity, and interest) (Rush *et al*, 2000, 2003b; Trivedi *et al*, 2004b). For three domains (sleep, psychomotor, appetite/weight disturbances), multiple questions are used (eg early, middle, late insomnia, and hypersomnia to evaluate sleep disturbance). For the sleep domain, the highest score on any of the four sleep items; for the appetite/weight domain, the highest score on any one of the four items, and for psychomotor changes, the highest score on either one of the two items were chosen to represent the domain. Each symptom item is scored on a 0–3 scale. The $QIDS\text{-}SR_{16}$ total score is computed by adding scores for the nine criterion domains. The $QIDS\text{-}SR_{16}$ total score ranges from 0 to 27. The PGI-I provided a seven-point rating with 1 designating 'very much improved' and 7 designating 'very much worse.'

### Statistical Methods

For these analyses, response was defined *a priori* as a $\geqslant 50\%$ reduction in baseline total score for each scale ($IDS\text{-}SR_{30}$, $QIDS\text{-}SR_{16}$, and $HDRS_{24}$) or a score of 1 or 2 for the PGI-I. For the $HDRS_{24}$, a total score $\leqslant 8$ was used to indicate remission, since it was used by Keller *et al* (2000). The remission thresholds for the $IDS\text{-}SR_{30}$ of $\leqslant 14$ and $QIDS\text{-}SR_{16}$ of $\leqslant 5$ were established using Item Response Theory (IRT) (Orlando *et al*, 2000) analysis. These thresholds were chosen as they correspond to a score of $\leqslant 7$ on the $HDRS_{17}$ (Rush *et al*, 2003b). These thresholds were verified (unpublished data) in a second dataset described by Trivedi *et al* (2001). The strength of agreement between each self-report ($IDS\text{-}SR_{30}$, $QIDS\text{-}SR_{16}$, and PGI-I) and the $HDRS_{24}$ was assessed for all patients and completers by the intraclass correlation coefficient (Shrout and Fleiss, 1979).

Response and remission rates were determined at exit for all patients (intent-to-treat) (ITT) and for completers. Exit status was assessed whenever the subject left the study (ITT sample) or at completion of the full 12 weeks of treatment (completer sample). The groups were compared by $\chi^2$ tests. Kappa statistics were used to measure agreement between response or remission groups as defined by each instrument and the $HDRS_{24}$. Change from baseline scores was compared between groups by *t*-test at each week. Patients were divided into response/remission groups based on $HDRS_{24}$ exit scores, and differences in $IDS\text{-}SR_{30}$, $QIDS\text{-}SR_{16}$, and PGI-I score were compared for these groups by *t*-test.

A mixed effects linear model similar to that used by Keller *et al* (2000) compared the treatment groups using the self-reports ($IDS\text{-}SR_{30}$, $QIDS\text{-}SR_{16}$, and PGI-I) and the clinician rating ($HDRS_{24}$) in the ITT sample. Thus, the treatment groups were compared based on the rate of change in the outcome measure from baseline to week 4 and from week 4 to exit as two separate analyses. The models have a random intercept and slope and terms for treatment group, site, time, and treatment $\times$ time interaction.

Evaluable samples had to include completed baseline and exit measurements (IDS-$SR_{30}$, PGI-I, and $HDRS_{24}$). Therefore, our sample of 602 subjects differs slightly from the Keller *et al* (2000) modified ITT sample of 656 subjects with postrandomization $HDRS_{24}$ scores (due to missing baseline PGI-I ($n=1$), baseline IDS-$SR_{30}$ ($n=27$); missing exit IDS-$SR_{30}$ ($n=26$)), although all of the Keller *et al* (2000) findings were confirmed in this sample ($n=602$).

Effect sizes were computed for $HDRS_{24}$, IDS-$SR_{30}$, and QIDS-$SR_{16}$ total scores as well as each item of the $HDRS_{24}$ and the IDS-$SR_{30}$ and for each domain of the QIDS-$SR_{16}$. Sensitivity to change was assessed by computing the effect size for change over time within each treatment group for total scores and for each item or domain (computed as mean change (baseline minus exit) divided by the standard deviation of the mean change). The ability of total scores and each item/domain to distinguish among treatment groups was assessed by computing the effect size of the treatment group effect (computed as the proportion of variance accounted for by treatment group membership as determined from an analysis of variance of baseline to exit change ($\omega 2$, Hays, 1988). The 95 percent confidence intervals (CI) were computed for within and between group total score effect sizes using the distribution of effect sizes generated from 5000 bootstrap samples (Davison and Hinkley, 1997).

## RESULTS

### Completion Rates

Of the 4788 actual visits, 122 visits (2.5%) were missing the IDS-$SR_{30}$, while 103 (2.2%) were missing the $HDRS_{24}$ rating. The PGI-I (which is not assessed at baseline) was missing for 132 out of 4186 post baseline visits (3.2%).

### Intraclass Correlation Coefficients and Coefficients of Variation

Intraclass correlation coefficients (ICC) were calculated for each of the three self-reports in relation to the clinician-rated $HDRS_{24}$ (Table 1) using all subjects ($n=602$) at exit and all completers ($n=490$) exit data. These ICCs are largely within the excellent ($\geqslant 0.75$) range (Fleiss, 1986).

The coefficients of variation (CV) were calculated for each scale using exit scores across all patients in all treatment groups. The CV (mean $\div$ standard deviation) were 1.46 ($HDRS_{24}$), 1.53 (IDS-$SR_{30}$), 1.52 (QIDS-$SR_{16}$), and 1.97 (PGI-I) indicating that for the IDS-$SR_{30}$ and QIDS-$SR_{16}$, there is no clinically significant difference in variability as compared to the $HDRS_{24}$. The PGI-I shows numerically higher CV suggestive of greater clinical variability. Note that the PGI-I generally has lower agreement with the $HDRS_{24}$ for remission at exit (Table 2 $\kappa$'s).

### Response and Remission Rates

A total of 602 patients were evaluable using the IDS-$SR_{30}$ (or QIDS-$SR_{16}$) and the PGI-I. Table 2 shows the response and remission rates for the ITT ($n=602$) and completer ($n=490$) samples using the IDS-$SR_{30}$, QIDS-$SR_{16}$, PGI-I, and $HDRS_{24}$. Each of the four ratings, in both the ITT and completer samples, revealed significantly higher response rates for the combination groups as compared to nefazodone alone or CBASP alone (all $\chi^2 \geqslant 27.8$; all $p<0.0001$). Response rates did not differ between the two monotherapies based on any of the three self-reports or on the $HDRS_{24}$ for both the ITT or the completer samples (all $\chi^2 \leqslant 1.6$; all $p \geqslant 0.20$). The $\kappa$'s for agreement between each instrument and the $HDRS_{24}$ were substantial (range = 0.61–0.80) (Landis and Koch, 1977), except for completers in the combination group, where the $\kappa$'s were moderate (range = 0.41–0.60) for the IDS-$SR_{30}$ and QIDS-$SR_{16}$.

Remission rates followed the same pattern (Table 3). The combination treatment was associated with higher remission rates based on the IDS-$SR_{30}$, QIDS-$SR_{16}$, and $HDRS_{24}$ (all $\chi^2 \geqslant 18.4$, all $p \leqslant 0.0001$) but not on the PGI-I (both $\chi^2 \leqslant 1.09$, $p \geqslant 0.58$) as compared to either monotherapy. Remission rates did not differentiate the two monotherapies for either the ITT or completer samples for any of the four outcome measures (all $\chi^2 \leqslant 0.47$, all $p \geqslant 0.49$).

Figure 1a shows the baseline and week-by-week total $HDRS_{24}$ scores for each treatment group. Figures 1b, c, and

**Table 1** Intraclass Correlations for Three Self-Report Measures Compared to the Clinician-Rated $HDRS_{24}$ by Treatment Group

| Sample | IDS-$SR_{30}$ | QIDS-$SR_{16}$ | PGI-I |
|---|---|---|---|
| *All patients ($n=602$)* | | | |
| Nefazodone ($n=196$) | 0.88 | 0.84 | 0.78 |
| CBASP ($n=199$) | 0.86 | 0.85 | 0.74 |
| Combination ($n=207$) | 0.82 | 0.79 | 0.74 |
| *Completers ($n=490$)* | | | |
| Nefazodone ($n=154$) | 0.89 | 0.84 | 0.80 |
| CBASP ($n=163$) | 0.88 | 0.86 | 0.77 |
| Combination ($n=173$) | 0.76 | 0.71 | 0.68 |

**Table 2** Response Rates for Three Self-Report Measures Compared to the Clinician-Rated $HDRS_{24}$ by Treatment Group

| Sample | IDS-$SR_{30}$[a] % ($\kappa$)[c] | QIDS-$SR_{16}$[a] % ($\kappa$) | PGI-I[b] % ($\kappa$) | $HDRS_{24}$[a] % |
|---|---|---|---|---|
| *All patients ($n=602$)* | | | | |
| Nefazodone ($n=196$) | 40.8 (0.72) | 45.4 (0.63) | 52.6 (0.74) | 47.4 |
| CBASP ($n=199$) | 41.2 (0.74) | 44.7 (0.77) | 48.2 (0.70) | 46.2 |
| Combination ($n=207$)[d] | 67.7 (0.67) | 67.6 (0.61) | 73.0 (0.72) | 75.8 |
| *Completers ($n=490$)* | | | | |
| Nefazodone ($n=154$) | 48.0 (0.72) | 52.6 (0.62) | 61.7 (0.72) | 55.8 |
| CBASP ($n=163$) | 46.0 (0.72) | 50.3 (0.75) | 54.6 (0.69) | 52.8 |
| Combination ($n=173$)[d] | 76.3 (0.54) | 75.7 (0.49) | 82.7 (0.62) | 85.0 |

[a]Response defined as a baseline to exit improvement $\geqslant$50%.
[b]Response defined as a PGI-I score of 1 or 2.
[c]$\kappa$ for agreement with $HDRS_{24}$ response status.
[d]The effect of the combination cell significantly exceeded the effect of either monotherapy ($p<0.0001$) for all outcome measures (all $\chi^2 \geqslant 27.8$).

**Table 3** Remission Rates for Three Self-Report Measures Compared to the Clinician-Rated $HDRS_{24}$ by Treatment Group

| Sample | $IDS\text{-}SR_{30}$[a] % ($\kappa$)[b] | $QIDS\text{-}SR_{16}$[a] % ($\kappa$) | PGI-I[a] % ($\kappa$) | $HDRS_{24}$[a] % |
|---|---|---|---|---|
| *All patients ($n=602$)* | | | | |
| Nefazodone ($n=196$) | 31.6 (0.66) | 27.6 (0.62) | 21.9 (0.64) | 29.6 |
| CBASP ($n=199$) | 31.7 (0.73) | 29.6 (0.71) | 21.1 (0.52) | 32.2 |
| Combination ($n=207$)[c] | 51.7 (0.60) | 50.2 (0.54) | 25.1 (0.40) | 49.3 |
| *Completers ($n=490$)* | | | | |
| Nefazodone ($n=154$) | 37.0 (0.66) | 33.1 (0.60) | 27.3 (0.66) | 35.7 |
| CBASP ($n=163$) | 36.2 (0.71) | 34.4 (0.69) | 23.9 (0.55) | 37.4 |
| Combination ($n=173$)[c] | 57.8 (0.55) | 56.1 (0.47) | 28.9 (0.34) | 56.6 |

[a]Remission defined as an exit score $\leqslant 14$ ($IDS\text{-}SR_{30}$), $\leqslant 5$ ($QIDS\text{-}SR_{16}$), $\leqslant 1$ (PGI-I), and $\leqslant 8$ ($HDRS_{24}$).
[b]$\kappa$ for agreement with $HDRS_{24}$ remission status.
[c]The effect of the combination cell significantly exceeded the effect of either monotherapy ($p=0.0001$) for all outcome measures except the PGI-I (all $\chi^2 \geqslant 18.4$).

d reveal very similar findings on a week-by-week basis using the $IDS\text{-}SR_{30}$, $QIDS\text{-}SR_{16}$, and PGI-I, respectively. In general, the same significant between-group differences in change from baseline on a week-by-week basis as results obtained with the $HDRS_{24}$ were found. However, the $IDS\text{-}SR_{30}$ did not distinguish CBASP from nefazodone or from the combination at week 1 and at week 6, and it did not distinguish nefazodone from CBASP at week 6. The $QIDS\text{-}SR_{16}$ revealed similar significant differences at the same visit occasions as the $HDRS_{24}$ except for lack of significant differences between CBASP and the other two treatments at week 1 and between nefazodone and the other two treatments at week 6. The PGI-I, while following a pattern similar to that found with the $HDRS_{24}$, was less consistent. The PGI-I did not differentiate the groups at week 1. Only trend differences were observed at week 2 (comparing the combination to CBASP alone) and at week 6 (comparing the combination to CBASP). However, exit findings with all three self-reports reflected the same significant between-group differences found with the $HDRS_{24}$ (significance test results not shown).

Table 4 provides results based on the $HDRS_{24}$ exit scores to define different groups (ie responder/nonresponder; remitter/nonremitter). All three self-reports distinguished $HDRS_{24}$ responders from $HDRS_{24}$ nonresponders, as well as $HDRS_{24}$ remitters from $HDRS_{24}$ nonremitters for each treatment group.

A further analysis compared $HDRS_{24}$ responders with residual symptoms (ie $\geqslant 50\%$ decrease from baseline but with an $HDRS_{24} > 8$) with $HDRS_{24}$ remitters (ie $HDRS_{24}$ score $\leqslant 8$ at exit). Results revealed mean $IDS\text{-}SR_{30}$ scores to be $18.2 \pm 7.8$ for $HDRS_{24}$ responders with residual symptoms *vs* $9.2 \pm 5.9$ for $HDRS_{24}$ remitters ($t=-11.0$, $df=191$, $p<0.0001$) at exit. For the $QIDS\text{-}SR_{16}$ results, $HDRS_{24}$ responders with residual symptoms averaged $7.1 \pm 3.2$ as compared to $3.8 \pm 2.6$ for $HDRS_{24}$ remitters ($t=-9.5$, $df=194$, $p<0.0001$). Finally, the mean score on the PGI-I for $HDRS_{24}$ responders with residual symptoms was $2.1 \pm 0.6$ as compared to $1.5 \pm 0.6$ for $HDRS_{24}$ remitters ($t=-7.2$, $df=340$, $p<0.0001$).

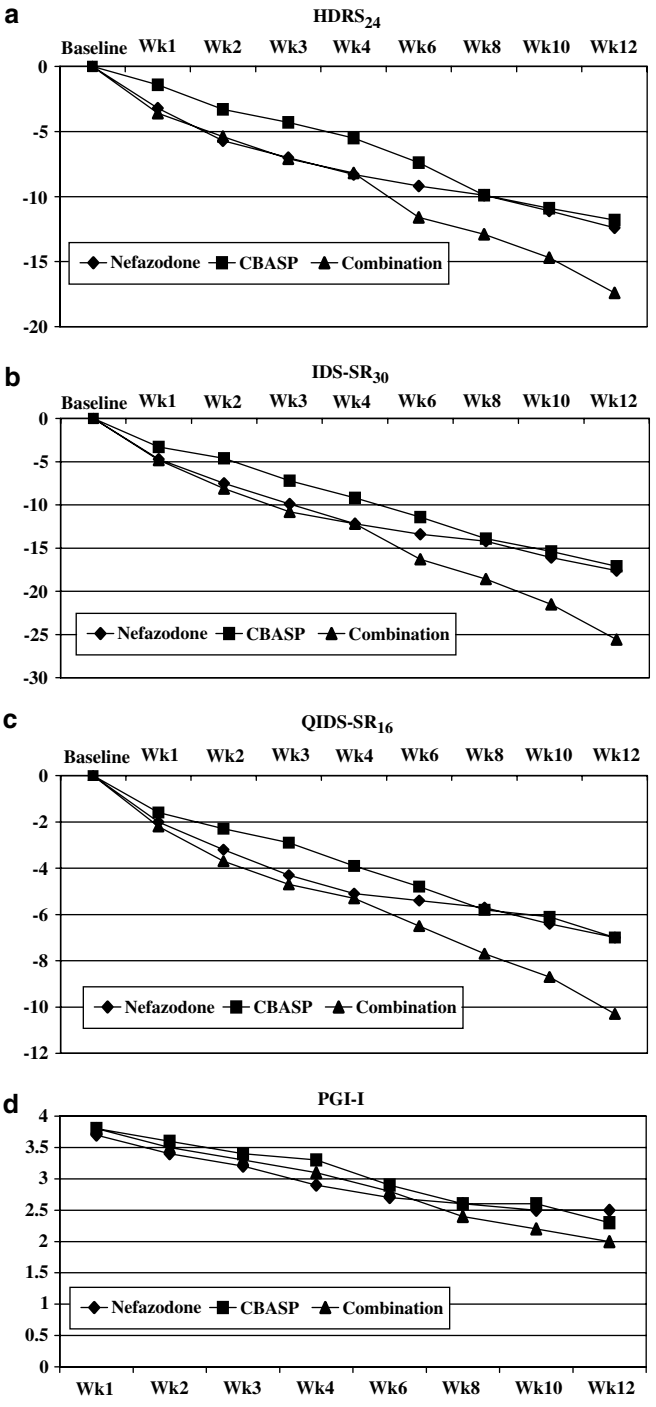


**Figure 1** (a) Mean change from baseline to week 12 in $HDRS_{24}$ total scores (observed case). (b) Mean change from baseline to week 12 in $IDS\text{-}SR_{30}$ total scores (observed case). (c) Mean change from baseline to week 12 in $QIDS\text{-}SR_{16}$ total scores (observed case). (d) Mean change from baseline to week 12 in PGI-I total scores (observed case).

We used a mixed effects model to parallel the original report based on the $HDRS_{24}$ (Keller *et al*, 2000), dividing the observation periods into the first 4 weeks and then the

**Table 4** Change from Baseline to Exit IDS-$SR_{30}$ and QIDS-$SR_{16}$ and PGI-I Scores by $HDRS_{24}$ Response Status at Exit by All Subjects and by Treatment Group[a]

| | $HDRS_{24}$ | | | |
|---|---|---|---|---|
| | **Responder[b]** | **Nonresponder** | **Remitter[c]** | **Nonremitter** |
| *IDS-$SR_{30}$* | | | | |
| All subjects | −26.7±10.1 ($n=342$) | −7.2±10.6 ($n=260$) | −29.6±9.5 ($n=224$) | −11.6±12.0 (n=378) |
| Nefazodone | −26.5±11.2 ($n=93$) | −7.5±10.4 ($n=103$) | −31.0±9.9 ($n=58$) | −10.4±11.2 (n=138) |
| CBASP | −25.3±9.6 ($n=92$) | −6.1±10.7 ($n=107$) | −28.2±8.7 ($n=64$) | −8.7±11.4 (n=135) |
| Combination | −27.6±9.7 ($n=157$) | −9.1±10.7 ($n=50$) | −29.7±9.7 ($n=102$) | −16.8±12.1 (n=105) |
| *QIDS-$SR_{16}$* | | | | |
| All subjects | −10.8±4.3 ($n=342$) | −2.9±4.6 ($n=260$) | −11.9±4.0 ($n=224$) | −4.7±5.1 ($n=378$) |
| Nefazodone | −10.5±4.9 ($n=93$) | −3.1±4.8 ($n=103$) | −12.3±4.4 ($n=58$) | −4.2±5.0 ($n=138$) |
| CBASP | −10.3±4.0 ($n=92$) | −2.4±4.4 ($n=107$) | −11.4±3.6 ($n=64$) | −3.5±4.8 ($n=135$) |
| Combination | −11.2±4.0 ($n=157$) | −3.8±4.7 ($n=50$) | −12.0±4.0 ($n=102$) | −6.9±5.0 ($n=105$) |
| *PGI-I* | | | | |
| All subjects | 1.7±0.7 ($n=342$) | 3.5±1.1 ($n=260$) | 1.5±0.6 ($n=224$) | 3.0±1.2 ($n=378$) |
| Nefazodone | 1.6±0.7 ($n=93$) | 3.5±1.2 ($n=103$) | 1.4±0.5 ($n=58$) | 3.2±1.3 ($n=138$) |
| CBASP | 1.7±0.7 ($n=92$) | 3.4±1.1 ($n=107$) | 1.6±0.7 ($n=64$) | 3.1±1.1 ($n=135$) |
| Combination | 1.8±0.7 ($n=157$) | 3.6±1.2 ($n=50$) | 1.6±0.7 ($n=102$) | 2.8±1.2 ($n=105$) |

[a]All differences between responders/nonresponders and between remitters/nonremitters were significant by *t*-test ($p<0.0001$).
[b]Responders achieved a $\geqslant$50% reduction in baseline $HDRS_{24}$ total score by exit.
[c]Remitters achieved an $HDRS_{24}$ total score $\leqslant$8 at exit.

**Table 5** A Comparison of Mixed Effects Model Results Based on $HDRS_{24}$, IDS-$SR_{30}$, QIDS-$SR_{16}$, or PGI-I as Outcome Measures

| | | ***p*-value*** | | | |
|---|---|---|---|---|---|
| **Time period** | **Comparison** | **$HDRS_{24}$** | **IDS-$SR_{30}$** | **QIDS-$SR_{16}$** | **PGI-I** |
| Baseline to week 4 ($n=602$) | Nefazodone *vs* CBASP | 0.0008 | 0.0064 | 0.0036 | 0.0003 |
| | Nefazodone *vs* combination | 0.84 | 0.72 | 0.57 | 0.22 |
| | Combination *vs* CBASP | 0.0005 | 0.0017 | 0.0004 | 0.011 |
| Week 4 to week 12 ($n=560$) | Nefazodone *vs* CBASP | 0.018 | 0.22 | 0.16 | <0.0001 |
| | Nefazodone *vs* combination | <0.0001 | <0.0001 | <0.0001 | <0.0001 |
| | Combination *vs* CBASP | 0.018 | <0.0001 | 0.0001 | 0.16 |

*A significant *p*-value indicates a difference in slope (baseline to week 4 or week 4 to exit) between treatment cells as defined by the mixed effects model.

subsequent 8 weeks of the 12-week trial. In this sample (baseline to week 4 ($n=602$) and week 4 to week 12 ($n=560$)), the $HDRS_{24}$ distinguished nefazodone from CBASP ($p<0.0008$) and distinguished the combination from CBASP ($p<0.0005$) over the first 4 weeks of the trial. The IDS-$SR_{30}$, QIDS-$SR_{16}$, and PGI-I mixed effects model analyses revealed similar statistically significant results. For the latter period (weeks 4–12), the $HDRS_{24}$, IDS-$SR_{30}$, and QIDS-$SR_{16}$ revealed greater efficacy for the combination than for either nefazodone alone or for CBASP alone (Table 5), while the PGI-I revealed greater efficacy for the combination *vs* nefazodone, but not for the combination *vs* CBASP alone.

Table 6 shows two different types of effect sizes for the $HDRS_{24}$ items, while Table 7 provides analogous information for the IDS-$SR_{30}$ items and QIDS-$SR_{16}$ domains. The first effect size is for change from baseline to exit within each treatment group expressed in standard deviation units. The second effect size is for the difference between treatment groups in change from baseline to exit expressed as a percent of variance explained. For the $HDRS_{24}$ items, item 1 (depressed mood) showed large within-group effects and the largest between treatment group effect of all the $HDRS_{24}$ items. Several $HDRS_{24}$ items showed very small within or between treatment group effects (eg items 9, 15, 17, 19, 20, and 21).

**Table 6** Effect Size (ES) for Baseline to Exit Change[a] and Treatment Group Effect in $HDRS_{24}$

| Item | Nefazodone ($n = 196$) | CBASP ($n = 199$) | Combination ($n = 207$) | Treatment group effect |
|---|---|---|---|---|
| 1—Depressed mood | 0.83 | 0.89 | 1.56 | 0.24 |
| 2—Guilt feelings | 0.76 | 0.74 | 1.23 | 0.17 |
| 3—Suicide | 0.52 | 0.42 | 0.75 | 0.10 |
| 4—Insomnia (early) | 0.49 | 0.29 | 0.65 | 0.14 |
| 5—Insomnia (middle) | 0.52 | 0.50 | 0.82 | 0.10 |
| 6—Insomnia (late) | 0.42 | 0.26 | 0.68 | 0.15 |
| 7—Work and activities | 0.96 | 0.84 | 1.44 | 0.18 |
| 8—Retardation | 0.26 | 0.42 | 0.53 | 0.10 |
| 9—Agitation | 0.23 | 0.28 | 0.30 | 0.00 |
| 10—Anxiety (psychic) | 0.66 | 0.73 | 1.09 | 0.16 |
| 11—Anxiety (somatic) | 0.48 | 0.52 | 0.67 | 0.04 |
| 12—Somatic symptoms (gastrointestinal) | 0.05 | 0.34 | 0.28 | 0.11 |
| 13—Somatic symptoms (general) | 0.70 | 0.69 | 1.10 | 0.14 |
| 14—Genital symptoms | 0.28 | 0.27 | 0.62 | 0.16 |
| 15—Hypochondriasis | 0.32 | 0.34 | 0.43 | 0.00 |
| 16—Weight loss | −0.13 | 0.02 | 0.06 | 0.06 |
| 17—Insight | 0.04 | 0.16 | 0.00 | 0.06 |
| 18—Diurnal variation (intensity) | 0.40 | 0.36 | 0.69 | 0.10 |
| 19—Depersonalization and derealization | 0.16 | 0.14 | 0.23 | 0.00 |
| 20—Paranoid symptoms | 0.15 | 0.25 | 0.09 | 0.07 |
| 21—Obsessional and compulsive symptoms | 0.07 | 0.20 | 0.16 | 0.00 |
| 22—Helpless | 0.80 | 0.55 | 1.11 | 0.16 |
| 23—Hopeless | 0.69 | 0.52 | 1.09 | 0.18 |
| 24—Worthless | 0.72 | 0.72 | 0.98 | 0.12 |
| Mean (SD) item ES | 0.43 (0.3) | 0.44 (0.2) | 0.69 (0.4) | 0.11 (0.1) |

[a]Within-group ES defined as (baseline mean−exit mean)/(SD of baseline−exit difference).

Considering the $IDS\text{-}SR_{30}$ items, item 5 (sad mood) had the largest within-group effects, and one of the largest between treatment group effects of all the $IDS\text{-}SR_{30}$ items. Hypersomnia (item 4) showed a very small treatment group effect for the $IDS\text{-}SR_{30}$, likely due to the smaller proportion of patients with this symptom at baseline.

Turning to the $QIDS\text{-}SR_{16}$, nine domains had effect sizes above 0.50, although the appetite/weight domain had a very small treatment group effect (ie substantial change occurred in all three groups, but that change did not differentiate among the three groups).

Table 8 shows that the $HDRS_{24}$, $IDS\text{-}SR_{30}$, and $QIDS\text{-}SR_{16}$ total scores all had similar within-group and between-group effect sizes. The $IDS\text{-}SR_{30}$ had the numerically largest effect size in the nefazodone and CBASP groups, while the $HDRS_{24}$ had the largest effect size in the combination group. The numerically somewhat larger within treatment group effect sizes achieved with the PGI-I were accompanied by the numerically smallest between-group effect size. These findings suggest that patients globally reported larger effects of any treatment as compared to the itemized reports. This larger effect, perhaps due to the global nature of the ratings, likely limited the capacity of the PGI-I ratings to detect between-group differences (ie small between-group effect size).

## DISCUSSION

Overall, these *post hoc* analyses based on self-reports confirmed the between-group differences reported by Keller *et al* (2000) using the $HDRS_{24}$ completed by highly trained, certified raters who were masked to treatment assignment. The $IDS\text{-}SR_{30}$ and the $QIDS\text{-}SR_{16}$ were generally as sensitive to both change over time and to differences between treatment cells as was the masked $HDRS_{24}$ rating. In addition, the $IDS\text{-}SR_{30}$ and $QIDS\text{-}SR_{16}$ confirmed response and remission rates obtained using the $HDRS_{24}$. Correspondence between the PGI-I and the $HDRS_{24}$ was far less impressive.

In addition, changes in $IDS\text{-}SR_{30}$ and $QIDS\text{-}SR_{16}$ total scores at each measurement occasion closely paralleled those obtained with the $HDRS_{24}$. PGI-I total scores were less consistent than changes in $IDS\text{-}SR_{30}$, $QIDS\text{-}SR_{16}$, or $HDRS_{24}$ total scores. When exit status was defined by the $HDRS_{24}$ (responders *vs* nonresponders, remitters *vs* nonremitters, and responders with residual symptoms *vs* remitters), the $IDS\text{-}SR_{30}$, $QIDS\text{-}SR_{16}$, and the PGI-I exit scores significantly differentiated these groups.

While the PGI-I, a global self-report, did not perform as well as the $IDS\text{-}SR_{30}$, $QIDS\text{-}SR_{16}$, or the $HDRS_{24}$, the major results of this study were still confirmed with the PGI-I.

**Table 7** Effect size (ES) for Baseline to Exit Change[a] and Treatment Group Effect in IDS-$SR_{30}$ Items and QIDS-$SR_{16}$ Domains (in bold)

| Item | Nefazodone ($n = 196$) | CBASP ($n = 199$) | Combination ($n = 207$) | Treatment group effect |
|---|---|---|---|---|
| 1—Onset insomnia | 0.70 | 0.54 | 0.89 | 0.14 |
| 2—Mid insomnia | 0.59 | 0.55 | 0.73 | 0.08 |
| 3—Morning insomnia | 0.51 | 0.33 | 0.70 | 0.15 |
| 4—Hypersomnia | 0.26 | 0.32 | 0.34 | 0.00 |
| **5—Sad mood** | **1.00** | **0.94** | **1.73** | **0.21** |
| 6—Irritable | 0.74 | 0.59 | 1.04 | 0.16 |
| 7—Anxious | 0.80 | 0.61 | 1.07 | 0.15 |
| 8—Reactivity of mood | 0.50 | 0.42 | 0.97 | 0.20 |
| 9—Mood variation | 0.18 | 0.26 | 0.42 | 0.08 |
| 10—Quality of mood | 0.70 | 0.74 | 1.07 | 0.18 |
| 11/12—Appetite change | 0.58 | 0.51 | 0.68 | 0.05 |
| 13/14—Weight change | 0.28 | 0.35 | 0.13 | 0.05 |
| **15—Concentration** | **0.65** | **0.64** | **1.10** | **0.20** |
| **16—Self outlook** | **0.67** | **0.62** | **1.00** | **0.15** |
| 17—Future outlook | 0.67 | 0.51 | 1.03 | 0.18 |
| **18—Suicidal ideation** | **0.73** | **0.53** | **0.77** | **0.12** |
| **19—Involvement** | **0.59** | **0.63** | **1.14** | **0.22** |
| **20—Energy** | **0.74** | **0.80** | **1.29** | **0.18** |
| 21—Enjoyment | 0.65 | 0.58 | 1.25 | 0.22 |
| 22—Sexual interest | 0.40 | 0.31 | 0.65 | 0.17 |
| 23—Psychomotor slowing | 0.58 | 0.59 | 0.90 | 0.13 |
| 24—Psychomotor agitation | 0.52 | 0.45 | 0.64 | 0.02 |
| 25—Somatic complaint | 0.47 | 0.48 | 0.76 | 0.09 |
| 26—Sympathetic arousal | 0.30 | 0.43 | 0.26 | 0.02 |
| 27—Panic | 0.44 | 0.45 | 0.60 | 0.03 |
| 28—Gastrointestinal | 0.30 | 0.31 | 0.48 | 0.04 |
| 29—Sensitivity | 0.78 | 0.92 | 1.18 | 0.10 |
| 30—Leaden paralysis | 0.80 | 0.77 | 0.90 | 0.06 |
| **Weight change (QIDS-$SR_{16}$) (max of items 11, 12, 13, 14)** | **0.52** | **0.54** | **0.57** | **0.00** |
| **Sleep disturbance (QIDS-$SR_{16}$) (max of items 1, 2, 3, 4)** | **0.70** | **0.61** | **0.91** | **0.12** |
| **Psychomotor changes (QIDS-$SR_{16}$) (max of items 23, 24)** | **0.61** | **0.59** | **1.00** | **0.14** |
| Mean (SD) IDS-$SR_{30}$ ES | 0.58 | 0.54 | 0.85 | 0.12 |
| Mean (SD) QIDS-$SR_{16}$ ES | 0.69 | 0.66 | 1.06 | 0.15 |

[a]Within-group ES defined as (baseline mean−exit mean)/(SD of baseline−exit difference).

**Table 8** Mean Effect Sizes (ESs) for Total Score Baseline to Exit Change and Treatment Group Effect

| | Nefazodone ES (95% CI) | CBASP ES (95% CI) | Combination ES (95% CI) | Treatment group effect ES (95% CI) |
|---|---|---|---|---|
| $HDRS_{24}$ total score | 1.11 (0.93–1.29) | 1.03 (0.88–1.21) | 1.87 (1.62–2.18) | 0.26 (0.18–0.33) |
| IDS-$SR_{30}$ total score | 1.15 (1.00–1.31) | 1.07 (0.91–1.26) | 1.83 (1.61–2.09) | 0.25 (0.18–0.32) |
| QIDS-$SR_{16}$ total score | 1.08 (0.93–1.24) | 1.05 (0.90–1.23) | 1.81 (1.58–2.08) | 0.24 (0.17–0.32) |
| PGI-I | 1.92 (1.77–2.14) | 2.11 (1.93–2.33) | 1.96 (1.78–2.20) | 0.15 (0.07–0.23) |

Perhaps the specific wording on each of the four (0–3) ratings for each IDS-$SR_{30}$/QIDS-$SR_{16}$ symptom item reduced between-subject (or within-subject over time) variability, which sharpened their capacity to reliably differentiate the three treatment groups. Global ratings such as the PGI-I (as opposed to anchored, itemized symptom ratings) by depressed patients may be more influenced by the well-known cognitive bias found in

symptomatic patients. Present results would recommend against use of the PGI-I when itemized clinician or self-reports are used in a trial, as the PGI-I was least useful in detecting between-group differences.

These results contradict the suggestion by Greenberg *et al* (1992) and others (Edwards *et al* 1984; Lambert *et al*, 1986; Murray, 1989) that clinical ratings and self-reports diverge in assessing outcomes across different treatments. However, given the absence of pill placebo in this trial, these findings do not fully address the concerns raised by Greenberg *et al* (1992) regarding placebo-controlled clinical trials (although CBASP was a nonmedication treatment cell). These results also address the issue raised by Duncan and Miller (2000) who questioned whether the use of patient self-reports in the Keller *et al* (2000) trial would have resulted in less robust outcomes based on treatment cell assignments.

The present results also reveal substantial agreement between the $HDRS_{24}$ total scores and those of the $IDS\text{-}SR_{30}$ and $QIDS\text{-}SR_{16}$, with intraclass correlations largely $\geqslant 0.8$. However, differences are to be noted as well. For example, the proportion of responders was generally 6–9 percentage points lower with the $IDS\text{-}SR_{30}$ or $QIDS\text{-}SR_{16}$ than for the $HDRS_{24}$, when a $\geqslant 50\%$ reduction in baseline severity at exit was used as a threshold for response for all three measures. Notably, remission rates were nearly identical for all three measures. These two findings suggest that some subjects who achieved an $HDRS_{24}$ response did not achieve an equivalent degree of benefit as viewed by the self-reports. Perhaps those who achieve an $HDRS_{24}$ response but who continue to have some symptoms (ie not achieve an $HDRS_{24}$ remission) may continue to suffer a negative cognitive bias and, therefore, may overrate their self-reported symptoms in some cases, such that a response by either self-report is not quite achieved. Alternatively, the $HDRS_{24}$ may overestimate response compared to the self-reports if for some patients the baseline $HDRS_{24}$ total ratings were inadvertently inflated (since a minimal $HDRS_{24}$ threshold was required for study entry). The inflation of baseline severity based on the $HDRS_{24}$, if present, would only affect response rates; it would not affect remission rates.

The relatively modest $\kappa$ values for response rates or poorer $\kappa$ values for remission rates deserve comment. Who is a responder by $HDRS_{24}$ is tightly tied to the baseline $HDRS_{24}$ total score. As noted above, there is a tendency in acute phase RCTs towards inflation in baseline severity scores.

The most likely explanation for the modest $\kappa$'s is that the two scales measure different enough constructs that agreement is modest. The differences between the scale items are in fact by design. Evidence for differences is suggested by the effect sizes found for each item/domain on each scale (Tables 6–8).

The item effect sizes obtained both for within treatment groups (baseline to exit) and across treatment groups (baseline to exit) deserve comment. If one considers both monotherapies only, and accepts an effect size of $\geqslant 0.5$ as moderate and meaningful (Cohen, 1988), only nine of 24 $HDRS_{24}$ items achieve this level of significance. The items include depressed mood, guilt feelings, middle insomnia, work and activities, psychic anxiety, somatic symptoms, and helplessness, hopelessness, and worthlessness. Notably, only six of the 17 items of the $HDRS_{17}$ achieve this threshold. Furthermore, for the $HDRS_{24}$ or the $HDRS_{17}$, only three core criterion symptoms (sad mood, guilt, and interest—or work and activities) have effect sizes of $\geqslant 0.5$, although middle insomnia (one of the four potential types of sleep disturbance) also has an effect size $\geqslant 0.5$.

All nine domains for the $QIDS\text{-}SR_{16}$ have a $\geqslant 0.5$ effect size for both monotherapies. For both the $HDRS_{24}$ and the $QIDS\text{-}SR_{16}$, some domains have similar effect sizes (eg sad mood). On the other hand, some important DSM-IV criterion symptom domains (eg suicide, psychomotor changes, sleep disturbances) have greater effect sizes when measured by the $QIDS\text{-}SR_{16}$ than when measured by the $HDRS_{24}$.

Turning to the $IDS\text{-}SR_{30}$, using items with effect sizes $\geqslant 0.5$ for both monotherapies as a threshold, the following items did not achieve this level of significance: early morning insomnia, hypersomnia, mood reactivity, diurnal variation, weight change, sexual interest, psychomotor agitation, somatic complaints, sympathetic nervous system arousal, panic attacks, and gastrointestinal complaints. Most of these items are not core criterion symptoms of a major depressive episode. Note that the effect sizes for each individual sleep item on the $IDS\text{-}SR_{30}$, while substantial (save for hypersomnia), are more modest than the effect size for the sleep disturbance domain on the $QIDS\text{-}SR_{16}$ (0.70, 0.61, 0.91) for each of the three treatment cells. These data suggest that all sleep items for the $IDS\text{-}SR_{30}$ or potentially for the $HDRS_{17}$ might be combined to form a single domain score, as is done on the $QIDS\text{-}SR_{16}$.

Finally, turning from item effect sizes to total score effect sizes, Table 8 shows the $HDRS_{24}$, $IDS\text{-}SR_{30}$, and $QIDS\text{-}SR_{16}$ to be equally sensitive in detecting baseline to exit change within each treatment group (note the degree of overlap among the confidence intervals within each treatment group). These three itemized self-report measures are also equally sensitive in distinguishing between treatment groups (again, note the overlapping confidence intervals in the last column of Table 8). The $QIDS\text{-}SR_{16}$ captures as much information about the differential effects of the three treatments as the $IDS\text{-}SR_{30}$ or $HDRS_{24}$.

The PGI-I produces larger within treatment group effect sizes than the other measures—with the largest effect in the CBASP group (2.11) unlike the other measures, which show the largest effect in the combination group. The PGI-I does the poorest job of separating the treatment groups. Perhaps global judgments made by depressed patients vary more widely across subjects, thereby reducing the chances of identifying between-group differences. Self-reports like the $IDS\text{-}SR_{30}$ or $QIDS\text{-}SR_{16}$ that provide clear anchors for responding to each item appear as useful as clinician ratings.

Is it actually so surprising that the $IDS\text{-}SR_{30}$ and $QIDS\text{-}SR_{16}$, which focus on specific depressive symptoms, correspond to standard clinician ratings such as the $HDRS_{24}$? In fact, most itemized clinician ratings actually rely on patients' self-reports to questions posed by the interviewer (eg suicidal thinking, nature of sleep, energy, guilt, etc.). The interviewer has the context of other depressed patients by which to gauge his/her rating, and the opportunity to clarify for patients (during the interview) the meaning of the question and the nature of the answer if patients have difficulty understanding the question or have

difficulty articulating the appropriate response. However, a few $HDRS_{24}$ items do require direct observation (eg psychomotor changes) and do not score subject reports of feeling agitated or slowed down. This difference could lead to increased sensitivity to psychomotor changes in self-reported measures, or at least lead to discrepancies between self-report and clinician ratings of these items.

The present results suggest that the $QIDS\text{-}SR_{16}$ and the $IDS\text{-}SR_{30}$ (and potentially other self-reports) with clear anchors for each item rated may provide sufficient consistency of responses across patients that results with clinical ratings are accurately portrayed by patient self-report, at least in depressed outpatients without cognitive impairment. Such self-reports may be a cost-effective substitute for more time-consuming clinician ratings. Obviously, very severely depressed psychotic or cognitively impaired depressed patients may be unable to provide accurate enough self-reported symptoms to allow for use of self-reports instead of clinician ratings.

It is notable that we were missing $IDS\text{-}SR_{30}$ measures for only 2.5% of visits, which was comparable to the missing rates (2.2%) for the $HDRS_{24}$ measures, while the latter was defined *a priori* as the primary outcome measure. Such a result suggests a high degree of acceptance of self-reports by both clinicians and patients.

Given these results, it is logical to ask whether the $IDS\text{-}SR_{30}$ should be replaced by the $QIDS\text{-}SR_{16}$. The $QIDS\text{-}SR_{16}$ is as, or more, sensitive to change as the $IDS\text{-}SR_{30}$. The 16 items assess nine domains, each of which has at least a moderate effect size. Furthermore, the overall effect sizes of the total scores are equivalent and as robust as the $HDRS_{17}$ total scores. Thus, to measure the nine criterion symptom domains (either in practice or in trials), the QIDS is entirely satisfactory, and the IDS is not needed.

The IDS, however, provides an accounting of all melancholic and atypical symptoms, and it measures common important associated symptoms (eg irritability, anxiety, and sympathetic nervous system arousal). The IDS is a useful initial measure to provide information as to depressive 'subtypes.' In addition, most of these noncriterion items also have substantial effect sizes, some of which also differentiate between treatment groups (Table 7) (eg sexual interest, reactivity of mood, psychomotor slowing, and future outlook). The present report also shows that the full $IDS\text{-}SR_{30}$ total score does, in fact, perform as well as the $HDRS_{24}$ or $QIDS\text{-}SR_{16}$ in spite of the wider range of symptoms being assessed. Thus, investigators or clinicians can select either scale as an acceptable, sensitive outcome measure depending on time constraints and their wish to assess only core criterion symptoms or a wider range of symptoms.

The current report includes the following limitations. First, the analyses are *post hoc*. However, all of these self-reports were included in the original protocol and data analysis plans as secondary or confirmatory analyses. Second, patients were not blind to the type of treatment they received (nefazodone, CBASP, or the combination). It is interesting, however, to note that the outcomes using nonblind, patient-itemized self-reports were virtually identical to outcomes obtained by highly trained, certified, and masked raters who completed the $HDRS_{24}$. Third, these findings may have limited generalizability as the patient sample was restricted to a largely Caucasian sample of outpatients with nonpsychotic MDD with various types of chronic courses. Fourth, there was no predetermined method by which the order of test administration was given. Fifth, while we do not believe that $HDRS_{24}$ raters reviewed any self-report data, we cannot guarantee that some may have had access to these data before conducting the $HDRS_{24}$ interview. No raters were aware of the data analyses to compare these ratings *post hoc*. Finally, the $QIDS\text{-}SR_{16}$ was derived from the items obtained with the $IDS\text{-}SR_{30}$ rather than from an independently administered measure.

These results clearly suggest that itemized self-reports may potentially substitute for clinician ratings, especially in large, multisite clinical trials with substantial sample sizes. Obtaining these self-reports could be facilitated by a telephone-based interview that elicits responses from the subject to each symptom item. These so-called Interactive Voice Response (IVR) systems have been used successfully in randomized clinical trials (Piette, 2000; Kobak *et al*, 1997, 1999), although studies using $IDS\text{-}SR_{30}$ or $QIDS\text{-}SR_{16}$ scores obtained via IVR have yet to be reported. The STAR*D trial (Fava *et al*, 2003; Rush *et al*, 2004) is comparing rater-acquired $HDRS_{17}$ and $IDS\text{-}C_{30}$ with an IVR-acquired $QIDS\text{-}SR_{16}$ to compare the utility, sensitivity to change, and psychometric properties of each type of rating.

The present findings are consistent with the idea of using a self-report in lieu of a clinician rating as a primary outcome, or to gauge treatment benefit in routine practice. To further evaluate the potential of self-reports in research, one needs a comparison between identical clinician and self-report measures in placebo-controlled trials, with an experimental drug, a standard drug, and a placebo to ascertain whether self-reports are as sensitive as clinician ratings in differentiating drug from placebo. Further studies to compare self-reports (obtained either by IVR or paper and pencil tests) and clinician ratings (both the Montgomery Asberg Depression Rating Scale and the Hamilton Rating Scale for Depression) are indicated in diverse populations of depressed patients and in epidemiologic samples.

## ACKNOWLEDGEMENTS

This research was supported in part by grants from Bristol-Myers Squibb Company to the 12 participating sites, by NIMH Grant #MH68851, and by the Betty Jo Hay Distinguished Chair, the Rosewood Corporation Chair in Biomedical Science, and by the Sara E and Charles M Seay Center for Basic Research in Psychiatry. We thank Fast Word Inc., Dallas, Texas for secretarial assistance in preparation of this manuscript.

## REFERENCES

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association: Washington, DC.

Bech P (1992). Symptoms and assessment of depression. In: Paykel ES (ed). *Handbook of Affective Disorders*. Guilford Press: New York. pp 3–14.

Beck AT, Rush AJ, Shaw BF, Emery G (1979). *Cognitive Therapy of Depression*. The Guilford Press: New York.

Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961). An inventory for measuring depression. *Arch Gen Psychiatry* **4**: 561–571.

Biggs MM, Shores-Wilson K, Rush AJ, Carmody TJ, Trivedi MH, Crismon ML *et al* (2000). A comparison of alternative assessments of depressive symptom severity: a pilot study. *Psychiatry Res* **96**: 269–279.

Carpenter LL, Yasmin S, Price LH (2002). A double-blind, placebo-controlled study of antidepressant augmentation with mirtazapine. *Biol Psychiatry* **51**: 183–188.

Carroll BJ, Feinberg M, Smouse PE, Rawson SG, Greden JF (1981). The Carroll Rating Scale for depression. Development, reliability and validation. *Br J Psychiatry* **138**: 194–200.

Cohen J (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. L. Erlbaum Associates: Hillsdale, NJ.

Corruble E, Legrand JM, Duret C, Charles G, Guelfi JD (1999a). IDS-C and IDS-SR: psychometric properties in depressed in-patients. *J Affect Disord* **56**: 95–101.

Corruble E, Legrand JM, Zvenigoroweki H, Duret C, Guelfi JD (1999b). Concordance between self-report and clinician's assessment of depression. *J Psychiatr Res* **33**: 457–465.

Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* **16**: 297–334.

Davison AC, Hinkley DV (1997). *Bootstrap Methods and their Application*. Cambridge University Press: Cambridge, UK.

Depression Guideline Panel (1993). *Clinical Practice Guideline. Number 5. Depression in Primary Care: Vol. 1. Detection and Diagnosis*, AHCPR Publication No. 93-0550 US Department of Health and Human Services, Agency for Health Care Policy and Research: Rockville, MD.

Domken M, Scott J, Kelly P (1994). What factors predict discrepancies between self and observer ratings of depression? *J Affect Disord* **31**: 253–259.

Duncan BL, Miller SD (2000). Nefazodone, psychotherapy, and their combination for chronic depression [correspondence]. *N Eng J Med* **343**: 1042.

Edwards BC, Lambert MJ, Moran PW, McCully T, Smith KC, Ellingson AG (1984). A meta-analytic comparison of the Beck Depression Inventory and the Hamilton Rating Scale for Depression as measures of treatment outcome. *Br J Clin Psychol* **23**: 93–99.

Fava M, Rush AJ, Trivedi MH, Nierenberg AA, Thase ME, Sackeim HA *et al* (2003). Background and rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study. *Psychiatr Clin North Am* **26**: 457–494.

First MB, Spitzer RL, Gibbon M, Williams JBW (1997). *Structured Clinical Interview for DSM-IV™ Axis I Disorders (SCID-I), Clinician Version*. American Psychiatric Publishing, Inc.: Washington, DC.

Fleiss JL (1986). *The Design and Analysis of Clinical Experiments*. John Wiley & Sons: New York, NY.

Gibbons RD, Clark DC, Kupfer DJ (1993). Exactly what does the Hamilton Depression Rating Scale measure? *J Psychiatr Res* **27**: 259–273.

Greenberg RP, Bornstein RF, Greenberg MD, Fisher S (1992). A meta-analysis of antidepressant outcome under 'blinder' conditions. *J Consult Clin Psychol* **60**: 664–669.

Grundy CT, Kunnen KM, Lambert MJ, Ashton JE, Tovey DR (1994). The Hamilton Rating Scale for Depression: one scale or many? *Clin Psychol Sci Pract* **1**: 197–205.

Gullion CM, Rush AJ (1998). Toward a generalizable model of symptoms in major depressive disorder. *Biol Psychiatry* **44**: 959–972.

Guy W (1976). *ECDEU Assessment Manual for Psychopharmacology*, Revised Edition. Department of Health, Education and Welfare Publication No. 76-338. US Government Printing Office: Washington, DC.

Hamilton M (1960). A rating scale for depression. *J Neurol Neurosurg Psychiatry* **23**: 56–62.

Hamilton M (1967). Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* **6**: 278–296.

Hays WL (1988). *Statistics*, 4th edn. Rinehart & Winston, Inc.: Fort Worth, TX.

Hughes JR, Krahn D (1985). Blindness and the validity of the double-blind procedure. *J Clin Psychopharmacol* **5**: 138–142.

IsHak WW, Burt T, Sederer LI (eds) (2002). *Outcome Measurement in Psychiatry. A Critical Review*. American Psychiatric Publishing, Inc.:Washington, DC. pp 439–441.

Jenkins C, Carmody TJ, Rush AJ (1998). Depression in radiation oncology patients: a preliminary evaluation. *J Affect Disord* **50**: 17–21.

Keller MB, McCullough JP, Klein DN, Arnow B, Dunner DL, Gelenberg AJ *et al* (2000). A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. *N Eng J Med* **342**: 1462–1470.

Kobak KA, Greist JH, Jefferson JW, Mundt JC, Katzelnick DJ (1999). Computerized assessment of depression and anxiety over the telephone using interative voice response. *MD Comput* **16**: 64–68.

Kobak KA, Tayler LH, Dottl SL, Greist JH, Jefferson JW, Burroughs D *et al* (1997). Computerized screening for psychiatric disorders in an outpatient community mental health clinic. *Psychiatr Serv* **48**: 1048–1057.

Lambert MJ, Hatch DR, Kingston MD, Edwards BC (1986). Zung, Beck, and Hamilton rating scales as measures of treatment outcome: A meta-analytic comparison. *J Consult Clin Psychol* **54**: 54–59.

Landis RJ, Koch GG (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**: 159–174.

Mazure C, Nelson JC, Price LH (1986). Reliability and validity of the symptoms of major depressive illness. *Arch Gen Psychiatry* **43**: 451–456.

McCullough Jr JP (2000). *Treatment for Chronic Depression. Cognitive Behavioral Analysis System of Psychotherapy (CBASP)*. Guilford Press: New York.

Miller IW, Bishop S, Norman WH, Maddever H (1985). The modified Hamilton Rating Scale for Depression: reliability and validity. *Psychiatry Res* **14**: 131–142.

Montgomery SA, Åsberg M (1979). A new depression scale designed to be sensitive to change. *Br J Psychiatry* **134**: 382–389.

Murray E (1989). Measurement issues in the evaluation of psychopharmacological therapy. In: Fisher S, Greenberg RP (eds). *The Limits of Biological Treatments for Psychological Distress: Comparisons with Psychotherapy and Placebo*. Erlbaum: Hillsdale, NJ. pp 39–68.

Orlando M, Sherbourne CD, Thissen D (2000). Summed-score linking using item response theory: application to depression measurement. *Psychol Assess* **12**: 354–359.

Piette JD (2000). Interactive voice response systems in the diagnosis and management of chronic disease. *Am J Manag Care* **6**: 817–827.

Prusoff BA, Klerman GL, Paykel ES (1972). Concordance between clinical assessments and patients' self-report in depression. *Arch Gen Psychiatry* **26**: 546–552.

Rabkin JG, McGrath P, Stewart JW, Harrison W, Markowitz JS, Quitkin F (1986). Follow-up of patients who improved during placebo washout. *J Clin Psychopharmacol* **6**: 274–278.

Rush AJ, Carmody T, Reimitz P-E (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.

Rush AJ, Crismon ML, Kashner TM, Toprac MG, Carmody TJ, Trivedi MH *et al* (2003a). Texas Medication Algorithm Project, Phase 3 (TMAP-3): rationale and study design. *J Clin Psychiatry* **64**: 357–369.

Rush AJ, Crismon ML, Toprac MG, Shon SS, Rago WV, Miller AL *et al* (1999a). Implementing guidelines and systems of care: experiences with the Texas Medication Algorithm Project (TMAP). *J Pract Psychiatry Behav Health* **5**: 75–86.

Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA *et al* (2004). Sequenced Treatment Alternatives to Relieve Depression (STAR*D): rationale and design. *Control Clin Trials* **25**: 119–142.

Rush AJ, Giles DE, Schlesser MA, Fulton CL, Weissenburger JE, Burns CT (1986). The Inventory for Depressive Symptomatology (IDS): preliminary findings. *Psychiatry Res* **18**: 65–87.

Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH (1996). The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychol Med* **26**: 477–486.

Rush AJ, Hiser W, Giles DE (1987). A comparison of self-reported *versus* clinician-rated symptoms in depression. *J Clin Psychiatry* **48**: 246–248.

Rush AJ, Rago WV, Crismon ML, Toprac MG, Shon SP, Suppes T *et al* (1999b). Medication treatment of the severely and persistently mentally ill: the Texas Medication Algorithm Project. *J Clin Psychiatry* **60**: 284–291.

Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN *et al* (2003b). The 16-item Quick Inventory of Depressive Symptomatology (QIDS) Clinician Rating (QIDS-C) and Self-Report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry* **54**: 573–583.

Snaith P (1993). What do depression rating scales measure? *Br J Psychiatry* **163**: 293–298.

Stallone F, Mendlewicz J, Fieve R (1975). Double-blind procedure: an assessment in a study of lithium prophylaxis. *Psychol Med* **5**: 78–82.

Shrout PE, Fleiss JL (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* **86**: 420–428.

Surís A, Kashner TM, Gillaspy Jr JA, Biggs M, Rush AJ (2001). Validation of the Inventory of Depressive Symptomatology (IDS) in cocaine dependent inmates. *J Offender Rehab* **32**: 15–30.

Tondo L, Burrai C, Scamonatti L, Weissenburger JE, Rush AJ (1988). A comparison between clinician-rated and self-reported depressive symptoms in Italian psychiatric patients. *Neuropsychobiology* **19**: 1–5.

Trivedi MH, Rush AJ, Crismon ML, Kashner TM, Toprac MG, Carmody TJ *et al* (2004a). The Texas Medication Algorithm Project (TMAP): clinical results for patients with major depressive disorder. *Arch Gen Psychiatry* **61**: 669–680.

Trivedi MH, Rush AJ, Ibrahim HM, Carmody TJ, Biggs MM, Suppes T *et al* (2004b). 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.

Trivedi MH, Rush AJ, Pan J-Y, Carmody TJ (2001). Which depressed patients repond to nefazodone and when? *J Clin Psychiatry* **62**: 158–163.

Yonkers KA, Samson J (2000). Mood disorders measures. In: American Psychiatric Association Task Force for the Handbook of Psychiatric Measures (ed). *Handbook of Psychiatric Measures*. American Psychiatric Association: Washington, DC. pp 515–548.

Zung WWK (1965). A self-rating depression scale. *Arch Gen Psychiatry* **12**: 63–70.

Zung WWK (1986). Zung Self-Rating Depression Scale and depression status inventory. In: Sartorius N, Ban TA (eds). *Assessment of Depression*. Springer: Berlin. pp 211–231.