Published in final edited form as:

Psychol Med. 2005 May ; 35(5): 693–704.

Multiple measures, methods, and moments: a factor-analytic investigation of change in depressive symptoms during acute-phase cognitive therapy for depression

JEFFREY R. VITTENGL^{1,*}, LEE ANNA CLARK², DOLORES KRAFT³, and ROBIN B. JARRETT³

¹Division of Social Science, Truman State University, Kirksville, MO, USA

²Department of Psychology, University of Iowa, Iowa City, IA, USA

³Department of Psychiatry, University of Texas Southwestern Medical Center, Dallas, TX, USA

Abstract

Background. Researchers and clinicians use a variety of measures to assess depressive symptoms, including clinician reports [for example the 17-item Hamilton Rating Scale for Depression (HRSD-17) and the 30-item Inventory for Depressive Symptomatology — Clinician Report (IDS-C-30)]; and patient reports [for example, the Beck Depression Inventory (BDI) and the 30-item Inventory for Depressive Symptomatology — Self-Report (IDS-SR-30)]. Although their concurrent convergent validity is well established, the degree to which these measures reflect the same pattern and level of change during treatment is unclear.

Method. We assessed depressive symptoms with two clinician (HRSD-17, IDS-C-30) and two patient (BDI, IDS-SR-30) reports on 14–15 occasions in 127 out-patients with recurrent major depressive disorder treated with 20 sessions of acute-phase cognitive therapy.

Results. Factor analyses revealed that time was the major source of variation in depressive symptom severity and change, with distinct ‘early’ and ‘late’ assessment factors, regardless of measure or rater (patient v. clinician). Average symptom severity decreased rapidly on all measures until mid-treatment, after which gradual decreases continued through the end of treatment. Effect sizes for pre- to post-treatment change were large for all measures ($d=1.9-2.2$), although there were significant differences among measures at some time-points.

Conclusions. The HRSD-17, BDI, IDS-C-30 and IDS-SR-30 reflect essentially the same symptom severity and change constructs during acute-phase cognitive therapy. Clinicians and researchers using one of these measures may estimate scores on the others with the provided common-factor conversions.

INTRODUCTION

Measuring depressive symptoms is a common, important task in research and clinical settings. Several depressive symptom severity scales have been developed with convergent validity well established at a single time-point (Beck *et al.* 1988). The purpose of the current analyses was to examine the inter-relations of four popular depressive symptom severity measures at multiple time-points before, during, and after acute-phase cognitive therapy for depression (Beck *et al.* 1979). We used a dataset from a randomized clinical trial on relapse prevention

* Address for correspondence: Dr Jeffrey R. Vittengl, Division of Social Science, Truman State University, 100 East Normal Street, Kirksville, MO 63501-4221, USA. (Email: vittengl@truman.edu) (Email: Robin.Jarrett@UTSouthwestern.edu)

DECLARATION OF INTEREST None.

(Jarrett *et al.* 2001), which included 127 out-patients with recurrent major depressive disorder assessed with the 17-item Hamilton Rating Scale for Depression (HRSD-17; Hamilton, 1960), the Beck Depression Inventory (BDI; Beck *et al.* 1961), and the 30-item self-report (IDS-SR-30) and clinician-rated (IDS-C-30) forms of the Inventory for Depressive Symptomatology (Rush *et al.* 1986, 1996) at 14-15 time-points. Important unanswered questions include whether these measures reflect the same depressive symptom construct over treatment time and whether these measures exhibit the same pattern and level of change. Negative answers would call into question the accumulation of knowledge across treatment studies using different depressive symptom measures, whereas affirmative answers would augment concurrent validity evidence and support the inter-change ability of these measures in clinical and research contexts. For example, a clinician could compare a patient's weekly progress on the IDSSR-30 with benchmarks from a treatment study using the HRSD-17.

The convergent validity of depressive symptom measures is usually estimated concurrently and found to be moderate to high. For example, the BDI correlates about 0.70-0.80 with the HRSD-17 (Beck *et al.* 1988; Clark & Watson, 1991); the IDS-C-30 correlates about 0.70 and 0.90 with the BDI and HRSD-17, and the IDSSR-30 correlates about 0.90 and 0.80 with the BDI and HRSD-17 respectively (Rush *et al.* 1986, 1996). In contrast, depressive symptoms may be assessed repeatedly, and retest reliability estimates are reported less frequently, perhaps because depressive symptom measures are thought to reflect an unstable state construct. Retest reliability estimates include 0.65-0.96 over 4-15 days for the HRSD-17 (Williams, 1988; Potts *et al.* 1990; Reynolds & Kobak, 1995), and 0.48-0.86 over 5-30 days for the BDI (Beck *et al.* 1988). With state measures, progressively lower retest reliabilities are expected for longer retest intervals, such as repeated assessments over treatment. However, the combination of moderately high concurrent validity and retest reliability is difficult to predict. Cross-time convergent validity may be lowered only by the measures' finite retest stability if they reflect the same construct, or may be substantially lower if the measures tap only partially overlapping constructs that change differentially over time.

For example, self- and clinician reports may reflect partly independent depressive symptom constructs (Steer *et al.* 1987; Gullion & Rush, 1998), and these constructs may change differently over treatment (e.g. patients and clinicians might attend to different change processes or have different expectations for change that influence their ratings). In addition to rater differences, however, measures also may vary in both item content and response scale, which potentially limit their convergence. Controlling variability in item content and response scale, the IDS-C-30 and IDS-SR-30 are fully parallel patient and clinician depressive symptom severity measures (Rush *et al.* 1986, 1996). This study aimed to clarify inter-relations of change among clinician and self-report measures of depressive symptom severity across acute-phase cognitive therapy.

Changes in depressive symptom levels also provide valuable information regarding how measures function in a treatment context. Mean symptom levels often decrease rapidly early in treatment, then more slowly through the remainder of treatment (Howard *et al.* 1986; Ilardi & Craighead, 1994; Rush *et al.* 1996). This nomothetic approach to characterizing change can be augmented by accounting for between patient variability in the magnitude of change through effect-size statistics such as Cohen's *d* (Cohen, 1988). Placing measures on a common effect-size metric also facilitates comparisons of the magnitude of change between measures. For example, the HRSD-17 demonstrates a reliably larger effect-size change than the BDI in treatment studies (Edwards *et al.* 1984; Lambert *et al.* 1986; Sayer *et al.* 1993); whereas the IDS-C-30 may change similarly to or slightly more than the IDS-SR-30 (Corruble *et al.* 1999; Biggs *et al.* 2000; Trivedi *et al.* 2004). The current dataset permits comparison of change in all four measures to help separate measure *versus* rater influences on the magnitude of reported change.

The current analyses extend past work with Jarrett and co-workers' (2001) treatment dataset (Clark *et al.* 2003 *a, b*; Vittengl *et al.* 2003, 2004 *a, b*, in press). We examine questions regarding stability and change in depressive symptoms across multiple methods, measures, and moments (i.e. time) using factor analysis. We expected the four depressive symptom measures to reflect the same construct and to change similarly over acute-phase cognitive therapy; that is, we hypothesized a single, large factor. Other plausible patterns include depressive symptom observations segregating by rater (i.e. two factors, one each for patient and clinician reports) or by measure (i.e. four factors for four measures). If the measures reflect the same construct and change similarly during treatment, we planned to provide inter-scale conversions to support accumulation of knowledge across treatment studies using different measures (e.g. Sechrest *et al.* 1996; Rush *et al.* 2003) by allowing researchers and clinicians using one measure to estimate scores on others.

METHOD

Participants

Participants were adult out-patients presenting with DSM-IV non-psychotic, recurrent, major depressive disorder (APA, 1994). Inclusion criteria included clear inter-depressive episode recovery (≥ 2 months of at least nearly normal functioning) and a score ≥ 16 on the HRSD-17. Exclusion criteria included concurrent medical disorders potentially accounting for depressive symptoms, organic mental disorders, psychotic disorders, active substance abuse or dependence, primary obsessive compulsive or eating disorders, borderline personality disorder, and inability or unwillingness to complete questionnaires or to comply with the treatment protocol. Participants were recruited through media, printed announcements, and self- and practitioner-referrals. They completed telephone screening ($n > 3500$) and diagnostic interviews ($n = 608$), and provided informed consent to enter the protocol ($n = 156$). More detail about participants, recruitment, inclusion, and exclusion criteria is available in Jarrett *et al.* (2001).

Acute-phase cognitive therapy

Acute-phase cognitive therapy (Beck *et al.* 1979) was conducted by five experienced therapists within a 12-week protocol, including 20 individual sessions (50-60 minutes) held twice weekly for 8 weeks, then weekly for 4 weeks. No pharmacotherapy was provided. Acute-phase cognitive therapy is designed to reduce depressive symptoms by eliciting thoughts associated with negative affect, and teaching patients to evaluate the validity of such thoughts through logical and empirical methods, to generate more realistic alternatives when negative thoughts are not supported, and to employ problem-solving skills when negative conclusions are warranted. More detail about this treatment, including therapists' high level of competence in providing cognitive therapy, is available in Jarrett *et al.* (2001).

Assessment strategy and timing

Two pre-treatment assessments were used to establish study eligibility and to render diagnoses. Patients presented at the Department of Psychiatry, University of Texas Southwestern Medical Center, Dallas, Texas, and completed the Structured Clinical Interview for DSM-III-R (SCID out-patient version; Spitzer *et al.* 1989), with supplemental interview questions to assess DSM-IV disorders across the lifespan and history and features of illness, as well as other clinician-rated measures described below. The final assessment was conducted by a doctoral-level diagnostician. In addition to the interview-based measures used to establish study eligibility, patients were evaluated at multiple time-points with a broad range of symptom and psychosocial measures. Assessments were completed at the same session as the eligibility determination (*Pre1*), at a follow-up assessment before beginning acute-phase cognitive therapy (*Pre2*), weekly during the 20-session protocol (*T1*—*T20*), and approximately 1 week

after *T20 (Post)*. Depressive symptoms were measured with two self-report and two clinician-rating scales.

Measures

HRSD-17—The HRSD-17 (Hamilton, 1960) is a widely used, 17-item, clinician rating scale to assess severity of depressive symptoms. Scores range from 0 to 52. Higher values represent greater symptoms. The scale has demonstrated good inter-rater reliability ($r = 0.85$; Clark & Watson, 1991) and internal consistency (alphas of 0.88 and 0.89 in two large clinic samples; Rush *et al.* 1996). In the current sample, alpha internal consistency was adequate (median 0.77, range 0.70-0.85) with the exception of the *Pre1* (0.22), *Pre2* (0.23), and *T1* (0.34) assessments. However, the early HRSD-17 scores correlated highly (0.69, 0.72, 0.72 respectively) with the other clinician report, the IDS-C-30, completed by the same rater. This consistency among measures completed by the same rater supports their reliability (e.g. as non-parallel alternative forms; Nunnally, 1978), so we retained these data points for analysis.

BDI—The BDI (Beck *et al.* 1961) is a very widely used, 21-item, self-report measure of depressive symptom severity. Scores range from 0 to 63. Higher values represent greater symptoms. Beck *et al.* (1988) reported an average internal consistency of 0.87. Similarly, in the current sample, alpha internal consistency was good to excellent (median 0.91, range 0.81-0.93).

IDS-30—This 30-item scale (Rush *et al.* 1986, 1996) has both self-report (IDS-SR-30) and clinician (IDS-C-30) versions to measure the severity of depressive symptoms. Thirty items are available but only one item in each of two item-pairs is used for a particular patient (increase or decrease in appetite and weight), yielding 28 scored items. Scores range from 0 to 84. Higher values represent greater symptoms. Rush *et al.* (1986) reported internal consistency reliabilities of 0.85 (IDS-SR-30) and 0.88 (IDS-C-30). In the current sample, alpha internal consistency was moderate to high for the IDS-SR-30 (median 0.90, range 0.72-0.91). The IDS-C-30 followed the same pattern as the HRSD-17, the other clinician rated measure: alpha internal consistency for the IDS-C-30 was adequate (median 0.87, range 0.78-0.91) with the exception of the *Pre1* (0.53), *Pre2* (0.55), and *T1* (0.61) assessments. As above, because the HRSD-17 and IDS-C-30 correlated highly during these assessments, we retained the IDS-C-30 data-points for analysis.

Estimation of missing data

Because many multivariate analysis techniques (e.g. factor analysis) exclude cases with missing data, we estimated a limited number of missing values to increase the power and generalizability of our results. First, we examined a frequency distribution of missing values. Because two measures (BDI, IDS-SR-30) were scheduled at 14 assessments and two (IDS-C-30, HRSD-17) at 15 assessments, up to 58 missing values were possible per patient. Five patients had one missing value, 16 patients had two missing values, no patients had three missing values, and 29 patients had four or more missing values. We elected to estimate missing values for each patient with two or fewer using the Markov Chain Monte Carlo method (Schafer, 1997; SAS Institute Inc., 1999), an iterative technique that uses observed means and co-variances. After estimation of missing values, the complete-data sample increased from 106 to 127. Although we present only analyses including estimated data, analyses without estimated data supported substantively equivalent conclusions.

Effect-size computations

We use Cohen's (1988) effect size d to characterize relative change in depressive symptom measures. We also compute an extension of d to individual cases, d_i , to facilitate statistical

tests of differences among the four measures in patients' degree of change. Effect size d may be computed as the standardized difference between dependent means. For example,

$$d = \frac{X_A - X_B}{\sqrt{S.D.^2_{X_A} + S.D.^2_{X_B} - 2r_{AB}S.D._{X_A}S.D._{X_B}}},$$

where A and B represent observations at two time-points (e.g. pre- and post-treatment) and r_{AB} is the retest correlation. We applied this same transformation to individual scores to compute

$$d_i = \frac{X_{A_i} - X_{B_i}}{\sqrt{S.D.^2_{X_A} + S.D.^2_{X_B} - 2r_{AB}S.D._{X_A}S.D._{X_B}}},$$

where i refers to an individual patient. The d_i scores have $M = d$ and $S.D. = 1.0$. We test differences between mean d_i scores with dependent t tests and analyses of variance.

RESULTS

Inter-relations of depression symptom severity measures across acute-phase cognitive therapy

We conducted principal axis factor analyses to examine inter-relations of the four depressive symptom severity measures over all available assessments (15 for clinician ratings, 14 for self-reports). We analyzed zero-order correlations to examine inter-relations of symptom severity, and separately analyzed partial correlations controlling *Pre1* scores to examine inter-relations in changes in severity. Both analyses revealed a large first factor and smaller but substantial second and third factors (proportion of common variance accounted for by factors 1-3=55.5%, 12.9%, 6.1%, and 58.3%, 10.4%, 5.4%, in the zero-order and partial correlation analyses respectively). The first two assessments (*Pre1-Pre2* and *Pre2-T1* in the zero-order and partial correlation analyses respectively) did not load strongly on the first unrotated factor (median 0.25, range 0.06-0.33) but loaded more strongly on the second factor (median 0.48, range 0.31-0.66). Moreover, the two- and three-factor solutions were readily interpretable. Consequently, we eliminated the one-factor solution from further consideration.

We varimax rotated the two-factor solutions to simple structure, and the factor loadings (shown in Table 1) clearly indicated that time was the critical variable. [Oblique promaxrotated factors were very similar to the orthogonal varimax-rotated factors in the two- and three-factor solutions (Tucker-Lewis congruency coefficients > 0.96).]

Regardless of measure or method, later assessments (approximately *T11-Post*) loaded highly only on the first factors, whereas early assessments (*Pre1* through approximately *T3* for zero order correlations or *T7* for partial correlations) loaded highly only on the second factors, with temporally intermediate assessments often splitting loadings between factors. These data indicate that early and late depressive symptom severity and change were distinct, further illustrated by selected bivariate correlations presented in Table 2. Specifically, retest correlations were moderate to high close in time (e.g. from *Pre1* to *T1*, and *T20* to *Post*) but low across longer periods of time (e.g. *Pre1* to *T11* and *Post*), especially early in treatment. Similarly, convergent validity of symptom severity level and change in depressive symptoms was moderate at *T1* and high by *T11*. Symptom severity level and change were more unstable early compared with later in treatment.

The three-factor solutions also indicated that time was the critical variable. The three factors corresponded roughly to late, middle, and early assessment periods. The first factor was marked most clearly by the *T15–Post* assessments, the second factor by the *T5–T13* assessments, and the third factor by *Pre1*, *Pre2*, *T1*, and *T3* assessments. However, most of the second factor's markers also loaded significantly on either the first or third factor. In contrast, the first and third factors (late and early periods respectively) were more distinct. (Loadings for the three-factor solutions are available from the corresponding author.) Thus, both two- and three-factor solutions revealed relatively clear 'early' and 'late' factors, with 'middle' assessments appearing either as cross-loadings or cross-over factors between them. We present the two-factor solution for its greater simplicity and clarity.

Depression symptom severity levels across acute-phase cognitive therapy

Whereas factor analyses clarified patterns of relative changes in depressive symptom measures, to understand the absolute changes better, we examined mean scores for each measure by assessment (see Fig. 1). The four depression symptom severity measures had nearly identical patterns of mean raw score change. The average decrease in symptoms from *Pre1* to *Post* assessments was marked, but the rate of change varied over time. Specifically, decreases were relatively large and generally statistically significant from assessment-to-assessment from *Pre1* to *T11*, after which change was substantially smaller and generally not statistically significant from assessment-to-assessment, although a trend for decreases in depressive symptom severity continued.

We next computed d_i effect-size scores separately for the four depressive symptom scales to characterize change between *Pre1* and each of the 14 subsequent assessments taking between patient variability into account. A multivariate analysis of variance of d_i scores revealed significant effects for measure, $F(3, 124) = 35.88, p < 0.01$, Wilks' lambda = 0.55, assessment, $F(12, 115) = 35.86, p < 0.01$, Wilks' lambda = 0.21, and the measurer $\times$ assessment interaction, $F(36, 91) = 20.32, p < 0.01$, Wilks' lambda = 0.11. The assessment effect reflected changes (primarily increases) in mean d_i over treatment time as depressive symptoms decreased. Collapsing across time, mean d_i varied significantly among the HRSD-17 (2.10), IDS-C-30 (1.78), IDS-SR-30 (1.74), and BDI (1.50), p 's < 0.01 , except for the non-significant ($p = 0.50$) difference between the IDS-C-30 and IDS-SR-30. The HRSD-17 reflected the most change in depression severity scores and the BDI the least, with the IDS-C-30 and IDS-SR-30 falling between these extremes.

The ordering of effect sizes among measures was mostly consistent from *T1–T19*. However, mean d_i for the clinician-rated measures (IDSC-30 and HRSD-17) peaked at *T18* (2.33 and 2.93 respectively) and decreased through *Post* (1.91 and 1.96 respectively); whereas mean d_i for the patient reports (IDS-SR-30 and BDI) increased slightly from *T18* (2.16 and 1.93 respectively) to *Post* (2.23 and 1.97 respectively), accounting for the measurer $\times$ assessment interaction. This pattern of effect sizes appears partly to reflect small changes in differences in raw scores between self- and clinician reports (see Fig. 1). However, when comparing the mean d_i 's, these small differences were amplified because between-patients variability in self-report measures decreased slightly from *T18* to *Post* (BDI S.D.'s = 8.4 and 7.7; IDS-SR-30 S.D.'s = 13.8 and 13.1 respectively), whereas variability in clinician reports increased somewhat (HRSD-17 S.D.'s = 5.8 and 7.3; IDS-C-30 S.D.'s = 11.6 and 13.0 respectively), and correlations from *Pre1* to *T18* and *Post* respectively, were relatively stable (maximum |change| in retest $r = 0.03$). Given equivalent retest correlations, the effect size is increased by larger differences in raw scores and decreased by larger variance.

Inter-conversion of depression symptom severity scales—The above analyses suggest that over acute-phase cognitive therapy, the BDI, HRSD-17, IDS-C-30, and IDS-SR-30

mark the essentially same depressive symptom severity and change constructs. Based on these findings, clinicians and researchers using any one of these four depressive symptom measures can estimate scores on the others. To compute concurrent inter scale conversions, we pooled data across assessments (127 patients $\times$ 14 assessments = 1778 observations; *Pre2* data available only for the HRSD-17 and IDS-C-30 were excluded) and submitted the four depression scales to a principal axis factor analysis. We extracted a one-factor solution to estimate the scales' reflection of the underlying depressive symptom severity construct. The factor accounted for nearly all of the common variance (99.9%), and the BDI (0.93), HRSD-17 (0.94), IDS-C-30 (0.96), and IDS-SR-30 (0.96) all loaded highly on it. [With confirmatory factor analysis, a non-stationary dynamic model with a first-order-autocorrelated factor (see Hershberger, 1998) fit the pooled data, goodness-of-fit index = 0.96, comparative fit index = 0.99, non-normed fit index = 0.97. This model indicates that the four depressive symptom measures mark the same construct, and that the construct has decreasing stability over longer periods of time, consistent with the exploratory factor analyses.]

A one-factor solution represented the pooled data well because time was captured in the patient-assessment unit of analysis — the four measures have moderate to high convergent validity at any given assessment time. We then derived a linear factor score and regressed it on each raw-score depression scale. With the resulting regression constants and coefficients, one may convert among any of the four depressive symptom measures through their common factor.

To check the reliability of the factor-analytic conversions, we examined agreement of scores estimated by this method with scores estimated through traditional bivariate regression and percentile (rank-order) methods. We computed intra-class correlations [ICC; Shrout & Fleiss, 1979, formula (2, 1)] among scores estimated with the factor-analytic, regression, and percentile methods for the 12 pairwise scale conversions (each of the four depressive symptom scales predicting the other three). ICCs of 0 and 1.0 mark no and perfect agreement respectively. The factor-analytic method agreed well with the bivariate regression (median ICC 0.99, minimum 0.97) and percentile (median ICC 0.99, minimum 0.96) methods.

Fig. 2 depicts the scales' linear relations to the underlying factor for visual conversions, and Table 3 presents equations for more precise conversions. For example, to convert an IDSSR-30 score of 25 to the BDI using Fig. 2, locate 25 on the vertical axis, trace right to the intersection with the IDS-SR-30 line, trace down to the intersection with the BDI line, then trace left to the vertical axis to see the estimated BDI score (approximately 16). Using Table 3, an HRSD-17 score of 15 converts to an IDS-C-30 score of $(0.49 + 2.13 \times 15) = 32.44$. An advantage of the factor-analytic method over traditional bivariate regression methods is that back conversions are equivalent within rounding error — an IDS-C-30 score of 32.44 converts back to an HRSD-17 score of $(-0.23 + 0.47 \times 32.44) = 15$.

DISCUSSION

The current multi-measure, multi-method, multi-occasion analyses indicate that four popular depressive symptom severity measures, the HRSD-17, BDI, IDS-C-30, and IDS-SR-30, mark essentially the same severity and change constructs during acute-phase cognitive therapy administered to adult out-patients with recurrent major depressive disorder. Factor analyses revealed that the inter-relations in both symptom severity and change reflect assessment time, rather than other structures such as rater (patient *versus* clinician) or measure. Further, analyses of mean severity and effect-size scores indicated that, for each of the four measures, depressive symptom severity decreased rapidly early in treatment and was more stable by mid-treatment (session 11-13) through a 1-week post-treatment follow-up. This pattern of results supports the cross-time convergent validity of the HRSD-17, BDI, IDS-C-30, and IDS-SR-30 in a treatment context.

Because nomothetic change was similar among depressive symptom measures, and they appeared to mark the same construct, we provide conversions among scales based on their common factor. Our factor-analytic conversion technique agreed well with less efficient bivariate methods (regression and percentile conversions) in this sample, and our derived conversion formulas are similar to and complement formulas based on different patient samples (Rush *et al.* 1986), as well as conceptually similar but computationally more complex item response theory methods (Orlando *et al.* 2000; Rush *et al.* 2003). Empirical conversions among measures help make findings more meaningful and generalizable in treatment outcome research (Sechrest *et al.* 1996; Rush *et al.* 2003) and support dissemination of research findings to empirical clinical practice. For example, a clinician using one of the measures (e.g. IDS-SR-30) could compare local patients' symptom levels to treatment matching and outcomes research using different measures (e.g. HRSD-17).

Our results reinforce the common clinical experience of rapid improvement early in acute-phase cognitive therapy followed by more gradual change. However, depression symptom severity and change formed quasi-independent factors early and late in treatment; relative differences among participants in depressive symptom severity and change were not stable over time. This finding may reflect partly the differentiation of treatment responders from non-responders. Consistent with the literature, the distinct early and late factors also mean that prediction of treatment outcomes from early depression severity or change should be poor.

Our finding of consistency in raw-score changes among measures was qualified by differences in effect sizes. The effect-size metric reflects not only raw-score change but also between-patient variability in change from the *Pre1* baseline. Averaged across assessments, the HRSD-17 changed most and the BDI least, consistent with past research comparing these measures (Lambert *et al.* 1986; Sayer *et al.* 1993). Our finding of intermediate sensitivity for the IDS-C-30 and IDS-SR-30 is novel and would be bolstered by replication. If replicated, parallel construction and empirical consistency between the IDS-SR-30 and IDS-C-30 helps rule out the rater (clinician *versus* patient) explanation for the difference in effect sizes between the HRSD-17 and BDI. Nevertheless, all effect sizes were 'large' by Cohen's (1988) benchmarks.

Because factor analyses indicate that the measures reflect the same depressive symptom severity and change constructs, varying emphases in symptom content may contribute to effect size differences. For example, the HRSD-17, IDS-C-30, and IDS-SR-30 tap somatic anxiety more than does the BDI; the IDS-C-30 and IDS-SR-30 assess symptoms of atypical depression (e.g. increased appetite and sleep, leaden paralysis, reactive mood, and rejection sensitivity) not measured on the HRSD-17 and BDI; and the BDI has more items reflecting self-blame than the other measures (Gullion & Rush, 1998).

However, the average differences in effect sizes were qualified during the last few assessments. Effect sizes for clinician reports (HRSD-17, IDS-C-30) decreased from *T18* to *Post* due to small increases in symptom levels and moderate increases in between-participant variability, whereas effect sizes for self-reports (BDI, IDS-SR-30) were steady. Many processes potentially contribute to this difference between clinicians' and patients' ratings. For example, treating clinicians may group patients informally into 'responder' and 'non-responder' categories as the end of treatment approaches, or patients may present depressive symptoms differently to independent evaluators *versus* treating clinicians. Replication of a late decrease in effect sizes for clinician ratings would support additional research to clarify this phenomenon. However, the current findings reinforce the validity of less costly self-report assessment (Biggs *et al.* 2000), especially regarding expected stability in effect sizes at the end of treatment.

It is important to note that the current results are based on a sample of adult out-patients with moderate to severe recurrent major depressive disorder who completed a course of acute-phase cognitive therapy. Our results, therefore, may not generalize well to all depressed patients, to patients with other subtypes of major depressive disorder, or to patients not completing acute-phase cognitive therapy or those receiving other treatments. Nevertheless, we note that recurrent depression is the most common subtype of major depressive disorder (Kessler, 2002), and 81% (127/156) of the sample consenting to acute-phase cognitive therapy contributed to our analyses. We believe, consequently, that the results have wide applicability, and that we give clinicians the tools needed to compare their patients' outcomes with those reported in clinical trials.

ACKNOWLEDGEMENTS

The clinical trial was conducted at the University of Texas Southwestern Medical Center at Dallas, Department of Psychiatry, in the Psychosocial Research and Depression Clinic directed by Dr Jarrett and was supported in part by grants MH-38238 and MH-01571 from the National Institute of Mental Health, Bethesda, MD (Dr Jarrett). Gratitude is expressed to our colleagues for contributing to this research: Jeanette Doyle, Greg Eaves, Paul Silver, Marjorie Woodruff, Bethany Hampton, Catherine Judd, Douglas Lisle, Regina Kinney, Maria Marwill-Magee, Andrew Clifford, Martin Schaffer and Rodger Kobes provided clinical support. Research support was provided by Barbara Foster, Michelle White, Edna Christian, Joseph Begue, Julie Lowe, Daisha Cipher, Patricia Green, Demetria Clinton and Paula Reese. Brian F. Shaw rated the cognitive therapists. We appreciate the administrative support of Eric J. Nestler (The Lou and Ellen McGinley Distinguished Chair in Psychiatric Research), Department of Psychiatry, University of Texas Southwestern Medical Center at Dallas.

REFERENCES

- APA, . Diagnostic and Statistical Manual of Mental Disorders. 4th edn. American Psychiatric Association; Washington, DC: 1994.
- Beck, AT.; Rush, AJ.; Shaw, BF.; Emery, G. Cognitive Therapy of Depression. Guilford Press; New York: 1979.
- Beck AT, Steer RA, Garbin MG. Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation. *Clinical Psychology Review* 1988;8:77–100.
- Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Archives of General Psychiatry* 1961;4:561–571. [PubMed: 13688369]
- Biggs MM, Shores-Wilson K, Rush AJ, Carmody TJ, Trivedi MH, Crimson ML, Toprac MG, Mason M. A comparison of alternative assessments of depressive symptom severity: a pilot study. *Psychiatry Research* 2000;96:269–279. [PubMed: 11084222]
- Clark LA, Vittengl JR, Kraft D, Jarrett RB. Separate personality traits from states to predict future depression. *Journal of Personality Disorders* 2003a;17:152–172. [PubMed: 12755328]
- Clark LA, Vittengl JR, Kraft D, Jarrett RB. Shared, not unique, components of personality and psychosocial functioning predict depression severity after acute-phase cognitive therapy. *Journal of Personality Disorders* 2003b;17:406–430. [PubMed: 14632375]
- Clark LA, Watson D. Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology* 1991;100:316–333. [PubMed: 1918611]
- Cohen, J. Statistical Power Analyses for the Behavioral Sciences. 2nd edn. Lawrence Erlbaum; Hillsdale, NJ: 1988.
- Corruble E, Legrand JM, Zvenigorowski H, Duret C, Guelfi JD. Concordance between self-report and clinician's assessment of depression. *Journal of Psychiatric Research* 1999;33:457–465. [PubMed: 10504014]
- Edwards BC, Lambert MJ, Moran PW, McCully T, Smith KC, Ellingson AG. A meta-analytic comparison of the Beck Depression Inventory and the Hamilton Rating Scale for Depression as measures of treatment outcome. *British Journal of Clinical Psychology* 1984;23:93–99. [PubMed: 6722384]
- Gullion CM, Rush AJ. Toward a generalizable model of symptoms in major depressive disorder. *Biological Psychiatry* 1998;44:959–972. [PubMed: 9821560]
- Hamilton M. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry* 1960;23:56–61.

- Hershberger, SL. Dynamic factor analysis. In: Marcoulides, GA., editor. *Modern Methods for Business Research*. Erlbaum; Mahwah, NJ: 1998. p. 217-249.
- Howard KI, Kopta MS, Krause MS, Orlinsky DE. The dose-effect relationship in psychotherapy. *American Psychologist* 1986;41:159–164. [PubMed: 3516036]
- Ilardi SS, Craighead WE. The role of nonspecific factors in cognitive-behavioral therapy for depression. *Clinical Psychology: Science and Practice* 1994;1:138–156.
- Jarrett RB, Kraft D, Doyle J, Foster BM, Eaves GG, Silver PC. Preventing recurrent depression using cognitive therapy with and without a continuation phase: a randomized clinical trial. *Archives of General Psychiatry* 2001;58:381–388. [PubMed: 11296099]
- Kessler, RC. Epidemiology of depression. In: Gotlib, IH.; Hammen, CL., editors. *Handbook of Depression*. Guilford; New York: 2002. p. 23-42.
- Lambert MJ, Hatch DR, Kingston MD, Edwards BC. Zung, Beck, and Hamilton rating scale measures of treatment outcome: a meta-analytic comparison. *Journal of Consulting and Clinical Psychology* 1986;54:54–59. [PubMed: 3958302]
- Nunnally, JC. *Psychometric Theory*. 2nd edn. McGraw Hill; New York: 1978.
- Orlando M, Sherbourne CD, Thissen D. Summed-score linking using item response theory: application to depression measurement. *Psychological Assessment* 2000;12:354–359. [PubMed: 11021160]
- Potts MK, Daniels M, Burnam MA, Wells KB. A structured interview version of the Hamilton Depression Rating Scale: evidence of reliability and versatility of administration. *Journal of Psychiatric Research* 1990;24:335–350. [PubMed: 2090831]
- Reynolds WM, Kobak KA. Reliability and validity of the Hamilton Depression Inventory: a paper-and-pencil version of the Hamilton Depression Rating Scale Clinical Interview. *Psychological Assessment* 1995;7:472–483.
- Rush AJ, Giles DE, Schlessner MA, Fulton CL, Weissenburger JE, Burns CT. The Inventory for Depressive Symptomatology (IDS): preliminary findings. *Psychiatry Research* 1986;18:65–87. [PubMed: 3737788]
- Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychological Medicine* 1996;26:477–486. [PubMed: 8733206]
- Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC, Ninan PT, Kornstein S, Manber R, Thase ME, Kocsis JH, Keller MB. 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. *Biological Psychiatry* 2003;54:573–583. [PubMed: 12946886]
- SAS Institute Inc.. *SAS/STAT User's Guide (version 8)*. SAS Institute Inc.; Cary, NC: 1999.
- Sayer NA, Sackeim HA, Moeller JR, Prudic J, Devanand DP, Coleman EA, Kiersky JE. The relations between observer-rating and self-report of depressive symptomatology. *Psychological Assessment* 1993;5:350–360.
- Schafer, JL. *Analysis of incomplete multivariate data*. Chapman & Hall; New York: 1997.
- Sechrest L, McKnight P, McKnight K. Calibration of measures for psychotherapy outcome studies. *American Psychologist* 1996;51:1065–1071. [PubMed: 8870543]
- Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 1979;86:420–428.
- Spitzer, RL.; Williams, JBW.; Gibbon, M.; First, MB. *Structured Clinical Interview for DSM-III-R-Out-patient Version (with Psychotic Screen)*. New York State Psychiatric Institute; New York: 1989.
- Steer RA, Beck AT, Riskind JH, Brown G. Relationships between the Beck Depression Inventory and the Hamilton Rating Scale for Depression in depressed outpatients. *Journal of Psychopathology and Behavioral Assessment* 1987;9:327–339.
- Trivedi MH, Rush AJ, Ibrahim HM, Carmody TJ, Biggs MM, Suppes T, Crimson ML, Shores-Wilson K, Toprac MG, Dennehy EB, Witte B, Kashner TM. 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. *Psychological Medicine* 2004;34:73–82. [PubMed: 14971628]

- Vittengl JR, Clark LA, Jarrett RB. Interpersonal problems, personality pathology, and social adjustment after cognitive therapy for depression. *Psychological Assessment* 2003;15:29–40. [PubMed: 12674722]
- Vittengl JR, Clark LA, Jarrett RB. Improvement in social-interpersonal functioning after cognitive therapy for recurrent depression. *Psychological Medicine* 2004a;34:643–657. [PubMed: 15099419]
- Vittengl JR, Clark LA, Jarrett RB. Self-directed affiliation and autonomy across acute and continuation phase cognitive therapy for recurrent depression. *Journal of Personality Assessment* 2004b;83:235–247. [PubMed: 15548462]
- Vittengl JR, Clark LA, Jarrett RB. Validity of sudden gains in acute-phase treatment of depression. *Journal of Consulting and Clinical Psychology*.
- Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry* 1988;45:742–747. [PubMed: 3395203]

FIG. 1. Mean raw scores on four depressive symptom measures by assessment. Significance of changes is depicted between adjacent assessments. (BDI, Beck Depression Inventory; HRSD-17, 17-item Hamilton Rating Scale for Depression; IDS-C-30, Inventory for Depressive Symptomatology — Clinician; IDS-SR-30, Inventory for Depressive Symptomatology — Self-Report; *Pre1* and *Pre2*, first and second pre-treatment assessments respectively; *T1*—*T20*, assessment at treatment sessions 1-20; *Post*, post-treatment assessment.)

FIG. 2. Estimated depression symptom severity scores predicted from their common factor, ± 1 standard error of prediction for an individual's score. (BDI, Beck Depression Inventory; HRSD-17, 17-item Hamilton Rating Scale for Depression; IDS-C-30, Inventory for Depressive Symptomatology — Clinician; IDS-SR-30, Inventory for Depressive Symptomatology — Self-Report.)

Table 1.Varimax-rotated loadings from cross-time factor analysis of depressive symptom measures ($n=127$)

Measure	Factor 1 = 'Late'		Factor 2 = 'Early'	
	Zero-order	Partial	Zero-order	Partial
HRSD-17 <i>Pre1</i>	0.07	-	0.51	-
BDI <i>Pre1</i>	0.00	-	0.67	-
IDS-C-30 <i>Pre1</i>	0.02	-	0.55	-
IDS-SR-30 <i>Pre1</i>	-0.01	-	0.71	-
HRSD-17 <i>Pre2</i>	0.00	-0.08	0.60	0.31
IDS-C-30 <i>Pre2</i>	0.07	0.01	0.62	0.39
HRSD-17 <i>T1</i>	0.01	-0.07	0.64	0.37
BDI <i>T1</i>	0.05	-0.02	0.82	0.56
IDS-C-30 <i>T1</i>	0.14	0.09	0.74	0.55
IDS-SR-30 <i>T1</i>	0.07	0.02	0.79	0.53
HRSD-17 <i>T3</i>	0.29	0.17	0.68	0.68
BDI <i>T3</i>	0.21	0.12	0.75	0.68
IDS-C-30 <i>T3</i>	0.38	0.28	0.70	0.69
IDS-SR-30 <i>T3</i>	0.26	0.19	0.77	0.66
HRSD-17 <i>T5</i>	0.39	0.27	0.51	0.67
BDI <i>T5</i>	0.37	0.29	0.63	0.65
IDS-C-30 <i>T5</i>	0.50	0.38	0.52	0.67
IDS-SR-30 <i>T5</i>	0.48	0.42	0.63	0.65
HRSD-17 <i>T7</i>	0.60	0.48	0.40	0.61
BDI <i>T7</i>	0.53	0.41	0.55	0.67
IDS-C-30 <i>T7</i>	0.64	0.51	0.35	0.63
IDS-SR-30 <i>T7</i>	0.61	0.50	0.52	0.66
HRSD-17 <i>T9</i>	0.71	0.61	0.39	0.55
BDI <i>T9</i>	0.70	0.65	0.48	0.55
IDS-C-30 <i>T9</i>	0.76	0.67	0.34	0.49
IDS-SR-30 <i>T9</i>	0.73	0.66	0.48	0.58
HRSD-17 <i>T11</i>	0.70	0.62	0.19	0.40
BDI <i>T11</i>	0.76	0.68	0.33	0.48
IDS-C-30 <i>T11</i>	0.79	0.71	0.16	0.41
IDS-SR-30 <i>T11</i>	0.79	0.71	0.33	0.50
HRSD-17 <i>T13</i>	0.74	0.67	0.23	0.38
BDI <i>T13</i>	0.76	0.69	0.36	0.47
IDS-C-30 <i>T13</i>	0.79	0.72	0.22	0.40
IDS-SR-30 <i>T13</i>	0.81	0.73	0.32	0.49
HRSD-17 <i>T15</i>	0.72	0.67	0.11	0.26
BDI <i>T15</i>	0.75	0.70	0.24	0.33
IDS-C-30 <i>T15</i>	0.80	0.75	0.10	0.29
IDS-SR-30 <i>T15</i>	0.81	0.77	0.19	0.29
HRSD-17 <i>T17</i>	0.84	0.83	0.12	0.19
BDI <i>T17</i>	0.87	0.85	0.23	0.29
IDS-C-30 <i>T17</i>	0.88	0.87	0.15	0.22
IDS-SR-30 <i>T17</i>	0.87	0.86	0.20	0.25
HRSD-17 <i>T18</i>	0.81	0.82	0.11	0.12
BDI <i>T18</i>	0.89	0.87	0.19	0.26
IDS-C-30 <i>T18</i>	0.85	0.85	0.09	0.13
IDS-SR-30 <i>T18</i>	0.88	0.88	0.13	0.18
HRSD-17 <i>T19</i>	0.81	0.83	0.14	0.13
BDI <i>T19</i>	0.86	0.88	0.16	0.14
IDS-C-30 <i>T19</i>	0.86	0.87	0.11	0.11
IDS-SR-30 <i>T19</i>	0.86	0.88	0.17	0.15
HRSD-17 <i>T20</i>	0.84	0.85	0.03	0.07
BDI <i>T20</i>	0.87	0.89	0.11	0.12
IDS-C-30 <i>T20</i>	0.89	0.90	0.04	0.09
IDS-SR-30 <i>T20</i>	0.88	0.89	0.11	0.14
HRSD-17 <i>Post</i>	0.79	0.82	0.01	0.00
BDI <i>Post</i>	0.88	0.90	0.07	0.08
IDS-C-30 <i>Post</i>	0.82	0.86	-0.01	-0.03
IDS-SR-30 <i>Post</i>	0.87	0.90	0.04	0.04

Loadings ≥ 0.30 in bold type.Factor analyses of zero-order and partial correlations (controlling *Pre1*) test inter-relations of symptom severity and change respectively.

HRSD-17, 17-item Hamilton Rating Scale for Depression; BDI, Beck Depression Inventory; IDS-C-30, Inventory for Depressive Symptomatology — Clinician Report; IDS-SR-30, Inventory for Depressive Symptomatology — Self-Report; *Pre1* and *Pre2*, first and second pre-treatment assessments respectively; *T1-T20*, assessment at treatment sessions 1-20; *Post*, post-treatment assessment.

Table 2.Selected retest reliability and concurrent convergent validity correlations ($n=127$)

Assessment	Depressive symptom measure			
	HRSD-17	BDI	IDS-C-30	IDS-SR-30
Retest zero-order r (partial r controlling $Pre1$)				
Pre1-T1	0.43 (-)	0.75 (-)	0.58 (-)	0.66 (-)
T1-T11	0.16 (0.13)	0.28 (0.16)	0.23 (0.26)	0.28 (0.21)
T1-Post	0.09 (0.03)	0.18 (0.11)	0.14 (0.07)	0.14 (0.09)
T11-Post	0.49 (0.49)	0.67 (0.66)	0.54 (0.54)	0.61 (0.60)
T20-Post	0.78 (0.77)	0.90 (0.89)	0.79 (0.79)	0.86 (0.86)
Median Convergent Zer-order r (Partial r controlling $Pre1$)				
Pre1	0.42 (-)	0.48 (-)	0.54 (-)	0.54 (-)
T1	0.61 (0.43)	0.64 (0.47)	0.73 (0.64)	0.77 (0.64)
T11	0.81 (0.81)	0.84 (0.86)	0.85 (0.86)	0.85 (0.86)
T20	0.87 (0.84)	0.88 (0.88)	0.91 (0.91)	0.91 (0.91)
Post	0.84 (0.84)	0.87 (0.87)	0.89 (0.89)	0.89 (0.89)

Values of $r \geq 0.18$ and 0.23, p values < 0.05 and 0.01, two-tailed respectively.

Medians computed from each measure's correlations with the other three measures.

HRSD-17, 17-item Hamilton Rating Scale for Depression; BDI, Beck Depression Inventory; IDS-C-30, Inventory for Depressive Symptomatology — Clinician Report; IDS-SR-30, Inventory for Depressive Symptomatology — Self-Report; $Pre1$ and $Pre2$, first and second pre-treatment assessments respectively ; $T1$ - $T20$, assessment at treatment sessions 1-20; $Post$, post-treatment assessment.

Table 3.

Formulas for pairwise conversions among four depressive symptom measures based on their common factor

Measure estimated	Measure known			
	BDI	HRSD-17	IDS-C-30	IDS-SR-30
BDI =	-	$-0.97 + 1.50 \text{ (HRSD-17)}$	$-0.80 + 0.79 \text{ (IDS-C-30)}$	$-1.32 + 0.70 \text{ (IDS-SR-30)}$
HRSD-17 =	$0.65 + 0.67 \text{ (BDI)}$	-	$0.11 + 0.53 \text{ (IDS-C-30)}$	$-0.23 + 0.47 \text{ (IDS-SR-30)}$
IDS-C-30 =	$1.01 + 1.26 \text{ (BDI)}$	$-0.21 + 1.89 \text{ (HRSD-17)}$	-	$-0.65 + 0.89 \text{ (IDS-SR-30)}$
IDS-SR-30 =	$1.87 + 1.42 \text{ (BDI)}$	$0.49 + 2.13 \text{ (HRSD-17)}$	$0.73 + 1.13 \text{ (IDS-C-30)}$	-

BDI, Beck Depression Inventory; HRSD-17, 17-item Hamilton Rating Scale for Depression; IDS-C-30, Inventory for Depressive Symptomatology — Clinician Report; IDS-SR-30, Inventory for Depressive Symptomatology — Self-Report.

Observed ranges: BDI=0-51; HRSD-17=0-30; IDS-C-30=0-52; IDS-SR-30=0-63. To convert an HRSD-17 score of 20 to the IDS-SR-30, for example, use the formula $\text{IDS-SR-30} = 0.49 + 2.13 \text{ (HRSD-17)}$; substitute the known HRSD-17 score in the equation: $0.49 + 2.13(20)$ and solve to yield $\text{IDS-SR-30} \approx 43$.