A Comparison of the QIDS-C_{16}, QIDS-SR_{16}, and the MADRS in an Adult Outpatient Clinical Sample

Ira H. Bernstein, PhD, A. John Rush, MD, Diane Stegman, RN, Laurie Macleod, RN,
Bradley Witte, and Madhukar H. Trivedi, MD

ABSTRACT

Background: This study compared the 16-item Clinician and Self-Report versions of the Quick Inventory of Depressive Symptomatology (QIDS-C_{16} and QIDS-SR_{16}) and the 10-item Montgomery-Asberg Depression Rating Scale (MADRS) in adult outpatients. The comparison was based on psychometric features and their performance in identifying those in a major depressive episode as defined by the Mini-International Neuropsychiatric Interview for *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition.

Methods: Of 278 consecutive outpatients, 181 were depressed. Classical test theory, factor analysis, and item response theory were used to evaluate the psychometric features and receiver operating characteristic (ROC) analyses.

FOCUS POINTS

- This study compared the Quick Inventory of Depressive Symptomatology (QIDS) to the Montgomery-Asberg Depression Rating Scale.
- The study sample consisted of adult outpatients on any medication, in treatment, or not in treatment, making them somewhat different from samples previously utilized in studies of the the QIDS.
- The study also compared clinical and self-report measures and performed test equating using item response theory methodology.
- Both classical and modern psychometrics were utilized to draw relevant comparisons.

Dr. Bernstein is professor in the Department of Clinical Sciences at the University of Texas Southwestern Medical Center in Dallas. Mr. Witte is data analyst in the Department of Psychiatry at UT Southwestern Medical Center. Dr. Rush is vice dean of the Duke-NUS Graduate Medical School in Singapore. Ms. Stegman retired as clinical research coordinator in the Department of Psychiatry at UT Southwestern Medical Center. Ms. Macleod is clinical research coordinator at UT-Tyler Medical Center. Dr. Trivedi is professor of psychiatry in the Department of Psychiatry at UT Southwestern Medical Center.

Faculty Disclosures: Dr. Bernstein owns stock in Bristol-Myers Squibb and Merck. Dr. Rush has been a consultant to Advanced Neuromodulation Systems, AstraZeneca, Best Practice Project Management, Bristol-Myers Squibb, Cyberonics, Forest Pharmaceuticals, GlaxoSmithKline, Magellan Health Services, Merck, Ono Pharma USA, Organon, Otsuka, Pamlab, Pfizer, and Transcept; is on the speaker's bureaus for Cyberonics, Forest, and Pfizer; has received research support from the National Institute of Mental Health (NIMH), and the Stanley Medical Research Institute; and has received royalty income from Guilford Publications and Healthcare Technology Systems. Ms. Stegman, Ms. Macleod, and Mr. Witte report no affiliation with or financial interest in any organization that may pose a conflict of interest. Dr. Trivedi has been a consultant to Abbott, Akzo, AstraZeneca, Bayer, Bristol-Myers Squibb, Cephalon, Cyberonics, Eli Lilly, Fabre-Kramer, Forest Pharmaceuticals, GlaxoSmithKline, Janssen, Johnson & Johnson, Meade Johnson, Neuronetics, Parke-Davis, Pfizer, Pharmacia & Upjohn, Sepracor, Solvay, VantagePoint, and Wyeth; has served on the speakers bureaus of Abdi Brahim, Akzo, Bristol-Myers Squibb, Cephalon, Cyberonics, Eli Lilly, Forest Pharmaceuticals, GlaxoSmithKline, Janssen, Pharmacia & Upjohn, Solvay, and Wyeth; and has received grant support from Bristol-Myers Squibb, Cephalon, Corcept, Cyberonics, Eli Lilly, Forest Pharmaceuticals, GlaxoSmithKline, Janssen, Merck, the NIMH, the National Alliance for Research in Schizophrenia and Depression, Novartis, Pfizer, Pharmacia & Upjohn, Predix, Solvay, and Wyeth.

Funding/Support: This research was funded in part by the NIMH, National Institutes of Health (MH-68852 to the University of Texas at Arlington, Ira H. Bernstein, PhD, PI; and MH-68851 to the University of Texas Southwestern Medical Center at Dallas, A. John Rush, MD, PI). The NIMH had no involvement in study design; in the collection, analysis, and interpretation of data; nor in the writing of the report or the decision to submit the paper for publication.

Submitted for publication: September 26, 2009; Accepted for publication: June 4, 2010.

Please direct all correspondence to: Ira H. Bernstein, PhD, Department of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9066; Tel: 214-648-9543, Fax: 214-648-3934; E-mail: Ira.Bernstein@UTSouthwestern.edu.

https://doi.org/10.1017/S1092852900000389 Published online by Cambridge University Press

Results: All three measures were unidimensional. All had acceptable reliability (coefficient α=.87 for MADRS$_{10}$, .82 for QIDS-C$_{16}$, and .80 for QIDS-SR$_{16}$). Test information function was higher for the MADRS (ie, it was most sensitive to individual differences in levels of depression). The MADRS and QIDS-C$_{16}$ slightly but consistently outperformed the QIDS-SR$_{16}$ in differentiating between depressed versus nondepressed patients.

Conclusion: All three measures have satisfactory psychometric properties and are valid screening tools for a major depressive episode.

CNS Spectr. 2010;15(7):458-468.

INTRODUCTION

Several measures, including the 10-item Montgomery-Asberg Depression Rating Scale (MADRS)[1] and the 16-item Clinician-Rated and Self-Report versions of the Quick Inventory of Depressive Symptomatology (QIDS-C$_{16}$ and QIDS-SR$_{16}$, respectively)[2,3] are available to evaluate the severity of depressive symptomatology. These measures are particularly useful for patients with major depressive disorder (MDD) as defined by the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition-Text Revision.[4] However, tests that usefully assess depressive symptom severity may not be optimal in differentiating patients in a major depressive episode (MDE) from patients not in an MDE as defined by the Mini-International Neuropsychiatric Interview (MINI)[5] or Structured Clinical Interview for Axis I Disorders.[6] In fact, the performance of these measures as a screening tool to identify patients in an MDE has not been widely investigated.[7,8] Furthermore, whether the self-report measure (the QIDS-SR$_{16}$) performs as well as the two clinician ratings for screening purposes, given the time efficiency of the self-report, is an important practical question.

These considerations led us to conduct an extensive series of evaluations of the various psychometric features of the MADRS and the QIDS-C$_{16}$ and QIDS-SR$_{16}$ in a large sample of outpatients attending a public sector psychiatric outpatient clinic.

METHODS

Participants

The sample was recruited from outpatients seeking care at the Psychiatric Outpatient Clinic at Parkland Hospital between April 2004 and August 2006. All participants provided written informed consent. The study was approved and overseen by the Institutional Review Board at the University of Texas Southwestern Medical Center at Dallas.

Eligible participants were 18–65 years of age. Excluded were persons with any disorder that affected their cognitive performance or ability to consent, judged clinically. This included psychotic diagnoses other than showing psychotic features in their depression. Participants could be taking any medications (both general medical and psychiatric) at the time of evaluation. They could be on any treatment or no treatment.

Assessments

Patients were interviewed using the MINI for the *DSM-IV* and *International Classification of Diseases*, Tenth edition,[5] conducted by experienced, trained research nurses to define an MDE. In addition, the MINI, MADRS$_{10}$, QIDS-C$_{16}$, and QIDS-SR$_{16}$ were administered in a completely randomized order so that any given test could precede or follow any other test by the same person who conducted the MINI.

Test Scoring

All scoring was done in the conventional manner. Each version of the QIDS was scored by selecting the maximum (most pathological) response to four sleep items, four weight/appetite items, and two restlessness/agitation items, thus converting these 10 individual items into three domains (sleep, appetite/weight, and psychomotor symptoms). The remaining items each defined the remaining six domains (total=nine domains).[9,10] The total MADRS score was based on responses to ten individual items.[1]

Statistical Methods

First, classical test theory (CTT) analyses were performed to generate the mean and item/total correlation (r_{it}) for each item or domain, the scale coefficient α, the scale mean, and scale standard deviation. Correlations among the three measures were computed.

https://doi.org/10.1017/S1092852900000389 Published online by Cambridge University Press

Next, each measure was subjected to component analysis and parallel analysis to infer dimensionality. Component analysis is a form of factor analysis that uses unit rather than estimated communalities. Parallel analysis[11-14] involves factoring random matrices having the same number of variables (nine for the two versions of the QIDS and 10 for the $MADRS_{10}$) and number of subjects (278) in the real data. The crossover point of the random and real data scree defines the dimensionality. Item response theory (IRT) analyses are most valid when applied to unidimensional scales (eg, Lord[15]).

The Samejima[16,17] IRT model for graded measures was applied to each scale to serve three purposes. First, item responses to the two versions of the QIDS were compared to evaluate differences between the clinical and self-report ratings on identical domains. The four response categories (0–3) for each item were divided into three dichotomies: normal (0) versus pathological (1, 2, or 3); normal or mildly pathological (0 or 1) versus moderately or severely pathological (2 or 3); and normal, mildly pathological, or moderately pathological (0, 1, or 2) versus severely pathological (3). Three boundary response functions were generated per domain/item corresponding to the three dichotomies. Each function relates the probability of choosing the more pathological category as a function of the magnitude of the trait in question, denoted θ in general but denoting overall depression in this paper. These functions are in the form of logistic ogives, which closely resemble cumulative normal distributions. The slope of each of these three curves is assumed to be the same. This common value is denoted $\underline{a}$, which describes the ability of the item (or domain) to discriminate levels of θ or, in this case, depression. The $\underline{a}$ value serves a similar role to the r_{it} of CTT with which it is usually highly correlated over domains/items even though it arises from a different model. The three intercepts are denoted b_0, b_1, and b_2 (b_i generically). The higher each intercept, the less likely the more pathological of the two alternatives represented by that particular dichotomization is chosen. A normal distribution scale is employed. Thus, if the b_0 parameter estimate is .0, half the population has responded with one of the pathological alternatives and the remaining half responds with the normal alternative.

The QIDS-C_{16} and QIDS-SR_{16} were compared in terms of their relative ability to discriminate levels of depression severity (values of $\underline{a}$) and levels of response (values of b_i). The strategy is first to fit a model in which the parameter estimates (eg, $\underline{a}$, b_0, b_1, and b_2) are allowed to vary freely (assume different values between groups and items/domains). A test statistic is estimated from the fit of this model known as a likelihood-ratio χ^2. Next, one or more parameters are constrained to equality between tests depending upon the hypothesis to be tested. For example, the nine $\underline{a}$ estimates for the QIDS-C_{16} can be equated to the nine $\underline{a}$ estimates for the QIDS-SR_{16}. This provides a second value of χ^2. The difference between the two may be tested for significance with degrees of freedom (df) equal to the number of parameters that have been constrained (nine in this case). If significant, it implies that the slopes are not the same for the two tests. This does not assume that the slopes within a given test equal, only that the corresponding slopes between tests are equal. Similar tests may be performed on the 27 paired values of b_i (nine domains x three values of b_i/domain) with (in this case) 27 df. If these omnibus tests on the values of $\underline{a}$ and b_i indicate the two tests differ, tests may also be performed on individual domains. These will respectively have 1 and 3 df although individual tests may be performed on each of the three values of b_i. Since this comparison involves tests given to the same individuals, both the omnibus slope and intercept tests would be nonsignificant if the tests were totally equivalent. Of course, it is of interest to see if, in fact, clinical judgments and self-report are equivalent.

Second, the test information function (TIF) was used to describe each test's overall sensitivity to individual differences in θ (depression) as a function level of depression. TIF provides a reliability-like measure, but it varies over levels of θ rather than being a constant like coefficient α.

The IRT analysis complements the CTT analysis. The latter furnishes relatively familiar statistics, eg, item/total correlations, but does not provide as convenient a way to compare items on the two versions of the QIDS and make other comparisons that IRT does. Similarly, IRT's TIF provides a more detailed description of how well a test discriminates among levels of a trait like depression than does the CTT's omnibus internal consistency measure, coefficient α.

https://doi.org/10.1017/S1092852900000389 Published online by Cambridge University Press

Third, depressed patients were contrasted with non-depressed patients as defined by the MINI, using the MADRS_{10}, QIDS-C_{16}, and QIDS-SR_{16}. Applying the Samejima model to the MADRS can produce up to six intercepts per item since the scale uses a seven-point rating for each item rather than the four-point rating for each item used in the QIDS-C_{16} and QIDS-SR_{16}. The number of intercepts is in general one less than the number of categories. However, the MADRS categories were reduced to five intercepts since few responses fell in the most extreme category; estimates derived from these small frequencies would be unstable. This could potentially happen with the QIDS-C_{16} and QIDS-SR_{16} but did not do so in this sample, in part due to the smaller number of categories. Moreover, because different patients populate the groups in and not in an MDE, there should be differences in the intercepts (values of b_i). Specifically, patients in an MDE should have lower threshold values (values of b_i) than patients not in an MDE because the former are more likely to choose the pathological alternative. No clear prediction can be made regarding possible differences in slopes (values of $\underline{a}$).

Several additional analyses examined the relation of the total scale scores for the three measures to identify patients in an MDE based on the MINI. The first analysis compared effect sizes (difference in mean score between those in and not in an MDE divided by the standard deviation of these differences). This analysis and the associated analysis of variance (ANOVA) main effect of group upon each scale are linear criteria that describe the ability of each scale to discriminate. These univariate evaluations ignore covariances among the three measures. They are equivalent to the univariate ANOVA $\underline{F}$-tests (also presented).

Second, univariate logistic regressions were conducted in which the scales were separately related to the log odds ratio (logits) of being classified as being in an MDE relative to those diagnosed as not being in an MDE. Associated with this were receiver operating characteristic (ROC) analyses, separately for the MADRS_{10}, QIDS-C_{16}, and QIDS-SR_{16}. Because both analyses utilize the log odds ratios of the probabilities rather than the probabilities themselves, these analyses are loglinear rather than linear.

Third, the contribution of each measure to the MANOVA discriminant axis was obtained. This is also linear, but it is multivariate since it assesses the ability of a given scale to increment the other two, holding the latter constant.

Finally, logistic regressions were conducted in which all three scales and pairs of scales were jointly entered as predictors to assess the ability of each scale to relate to the log odds of classification holding the two other scales constant. This is both loglinear and multivariate.

These various methods were used since it is quite possible for one measure to be most successful by one criterion and another measure to be most successful by a different criterion. Consistent findings across these methods provide more convincing evidence of the certainty of findings.

The final step was to equate test scores in this sample using methods previously described.[18-20]

RESULTS

Of the 291 enrollees, five did not complete the three measures of interest (QIDS-C_{16}, QIDS-SR_{16}, and MADRS_{10}), and eight were in partial remission, so they were excluded from analysis. Of the remaining 278, 181 were classified as having a current clinical depression (169 in a MDE and 13 with dysthymia disorder). Of those in a current clinical depression, 69 had melancholic features and 38 had psychotic features. Of the remaining 97 nondepressed patients, 84 had never had an MDE, and 13 had had a previous MDE but were currently in remission. A preliminary analysis indicated that these two subgroups of patients classified as nondepressed did not differ significantly in terms of mean QIDS-C_{16}, QIDS-SR_{16}, or MADRS scores. The total sample was 64.3% female with a mean age of 42.9±10.1 years and a range from 18–65 years. The total sample divided into 50.7% White, 36.7% African-American, 10.1% Hispanic, and 2.5% other. Their marital status was 19.1% married or cohabitating, 38.5% divorced or separated, 37.4% never married, and 5.0% widowed. Table 1 provides the demographic data. None of the tabled differences were statistically significant.

CTT Analysis

Table 2 shows the results of the CTT analyses for the QIDS-C_{16} and QIDS-SR_{16}. Table 3 contains comparable data for the MADRS items. Note that all three scales were at least moderately

https://doi.org/10.1017/S1092852900000389 Published online by Cambridge University Press

reliable in this sample. The two versions of the QIDS differed minimally.

The MADRS and the QIDS-C$_{16}$ correlated .89 over all patients; the MADRS and the QIDS-SR$_{16}$ correlated .78, and the two versions of the QIDS correlated .80.

Scale Dimensionality

Figures 1 and 2 contain the scree plots for the QIDS-C$_{16}$, QIDS-SR$_{16}$, and the MADRS, respectively. All three measures were unidimensional.

IRT Analysis of Differences Between the QIDS-C$_{16}$ and QIDS-SR$_{16}$

The overall test of slope differences between the two tests was nonsignificant (χ^2[9]=3.8), but the overall test of intercept differences was significant, (χ^2[27]=71.4; *P*<.0001). The sad mood, concentration/decision making, self-view, general interest, energy level, and restlessness/agitation groups of intercepts all differed at least at the .05 level. However, the direction of difference was not consistent. In 11 of these 18 comparisons (six domains x three intercepts/domain), the QIDS-SR$_{16}$ led to more frequent choice of the pathological category than the QIDS-C$_{16}$, but in seven cases the reverse was true. There was no apparent pattern to these differences.

TABLE 1.
Demographic Characteristics of the Total Sample, Those Not in a Current MDE, and Those in a Current MDE

		Current MDE?	
Variable (%)	***All N=278***	***No (n=97)***	***Yes (n=181)***
Female	64.3	68.2	59.6
White	50.7	47.4	52.5
Hispanic	10.1	14.4	7.7
African American	36.7	35.1	37.6
Other ethnicity	2.5	3.1	2.2
Married/cohabitating	19.1	13.4	22.1
Divorced	25.9	25.8	26.0
Separated	12.6	9.3	14.4
Never married	37.4	47.4	32.0
Widowed	5.0	4.1	5.5
Employed full time	10.8	9.3	11.7
Employed part time	6.1	7.2	5.6
Unemployed	80.1	81.4	79.4
Retired	1.4	1.0	1.7
Other employment status	1.4	1.0	1.7
Mean age	42.9	41.7	43.6
SD age	10.1	10.1	10.1
Mean years of education	12.0	12.0	12.1
SD years of education	2.8	2.6	2.9

MDE=major depressive episode; SD=standard deviation.

Bernstein IH, Rush AJ, Stegman D, Macleod L, Witte B, Trivedi MH. *CNS Spectr.* Vol 15, No 7. 2010.

Test Information Functions

Figure 3 contains the test information functions for the three tests. The larger value of coefficient α for the MADRS is reflected in its greater test information starting at an inferred depression score of -2, which is near the sample minimum. Differences between the two versions of the QIDS were negligible.

Item Differences Between Patients in an MDE and not in an MDE

As expected, there were large intercept differences because depressed patients chose more pathological alternatives than those not depressed. The values of χ^2 were 165.8, 168.4, and 331.2 and for the QIDS-C$_{16}$, QIDS-SR$_{16}$, and MADRS (df=27, 27, and 50, respectively; *P*<.0001). The depressed patients always chose the more pathological alternative; these intercept differences were significant beyond the .05 level on all items from all three tests. The differences were significant beyond the .0001 level in 23 of these 28 comparisons. In contrast, none of the overall slope differences was significant (χ^2=8.3, 10.4, and 7.1; df=9, 9, and 10, respectively) which means that the items/domains were equally discriminating between the depressed and not depressed.

Screening Validity Based on Effect Sizes and ANOVA

The effect sizes (the mean difference between depressed and not depressed patients divided by the pooled within-groups standard deviation) for the QIDS-C$_{16}$ and the MADRS were nearly equal at 1.53 and 1.51, respectively. The effect size for the QIDS-SR$_{16}$ was slightly smaller at 1.37. The corresponding values of $\underline{F}$(1,276) obtained from univariate ANOVAs were 144.04, 147.04, and 118.57 (*P*<.0001, df=1 and 276).

https://doi.org/10.1017/S1092852900000389 Published online by Cambridge University Press

TABLE 2.

QIDS-C$_{16}$ and QIDS-SR$_{16}$ Domain Means, Values of r_{it}, Scale Means, Scale SDs, and Coefficients α for all Patients and Patients in and Not in a Current MDE

	QIDS-C$_{16}$						QIDS-SR$_{16}$					
	All N=278		Not in MDE n=97		MDE n=181		All N=278		Not in MDE n=97		MDE n=181	
Domain	***Mean***	r_{it}	***Mean***	r_{it}	***Mean***	r_{it}	***Mean***	r_{it}	***Mean***	r_{it}	***Mean***	r_{it}
Sleep	2.47	.44	2.16	.44	2.63	.34	2.50	.39	2.18	.27	2.67	.29
Sad mood	1.47	.64	.69	.37	1.89	.50	1.54	.64	.78	.43	1.94	.53
Appetite	1.81	.39	1.41	.24	2.02	.34	1.92	.30	1.67	.32	2.05	.23
Concentration/decision making	1.38	.56	.89	.32	1.64	.50	1.33	.61	.87	.57	1.58	.50
Self view	1.01	.51	.43	.29	1.33	.39	1.01	.46	.54	.47	1.27	.32
Thoughts of death or suicide	.56	.60	.18	.34	.77	.57	.61	.52	.23	.33	.82	.49
General interest	1.10	.58	.54	.29	1.41	.51	1.20	.51	.61	.27	1.51	.43
Energy level	1.42	.62	.74	.34	1.78	.51	1.45	.58	.81	.43	1.79	.46
Restlessness/agitation	1.27	.45	1.08	.39	1.37	.45	1.51	.45	1.30	.55	1.62	.40
Scale mean	12.50		8.12		14.84		13.06		8.98		15.25	
Scale SD	5.44		3.69		4.74		5.46		4.42		4.66	
α	.82		.65		.77		.80		.72		.72	

QIDS-C$_{16}$=16-item Quick Inventory of Depressive Symptomatology–Clinician-Rated; QIDS-SR$_{16}$: 16-item Quick Inventory of Depressive Symptomatology–Self-Report; MDE=major depressive episode; SD=standard Deviation.

Bernstein IH, Rush AJ, Stegman D, Macleod L, Witte B, Trivedi MH. *CNS Spectr.* Vol 15, No 7. 2010.

TABLE 3.

MADRS Item Means, Values of r_{it}, Scale Mean, Scale SDs, and Coefficients α for all Patients, Patients Not in a Current MDE, and Patients in a Current MDE

	All N=278		Not in MDE n=97		MDE n=181	
MADRS Item	***Mean***	r_{it}	***Mean***	r_{it}	***Mean***	r_{it}
Apparent sadness	1.92	.68	1.04	.56	2.39	.57
Reported sadness	2.35	.77	1.04	.51	3.06	.69
Inner tension	2.23	.52	1.66	.39	2.53	.47
Reduced sleep	2.40	.46	1.67	.38	2.80	.38
Reduced appetite	1.31	.43	.75	.36	1.61	.37
Concentration difficulties	2.40	.57	1.48	.43	2.88	.47
Lassitude	2.37	.64	1.32	.42	2.93	.55
Inability to feel	1.99	.68	.93	.49	2.56	.58
Pessimistic thoughts	1.63	.63	.82	.43	2.06	.57
Suicidal thoughts	.85	.58	.18	.42	1.22	.50
Scale mean	19.45		10.90		24.03	
Scale SD	10.71		6.93		9.51	
α	.87		.76		.82	

MADRS=10-item Montgomery-Asberg Depression Rating Scale; SD=standard deviation; MDE=major depressive episode.

Bernstein IH, Rush AJ, Stegman D, Macleod L, Witte B, Trivedi MH. *CNS Spectr.* Vol 15, No 7. 2010.

https://doi.org/10.1017/S1092852900000389 Published online by Cambridge University Press

Screening Validity Based on Univariate Logistic Regression and ROC Analyses

The residual chi-square adjusting only for the proportion of depressed or not depressed patients (.65 versus .35, respectively) was 347.08 on 277 df, *P*<.01. Individually entering the $MADRS_{10}$, $QIDS\text{-}C_{16}$, and $QIDS\text{-}SR_{16}$ significantly reduced these by 119.9, 116.7, and 97.1 (*P*<.0001). The associated regression weights were .18, .35, and .29 (the ranking of the chi-square changes did not agree with the rankings of the regression weights because the standard errors which influenced the former were also unequal). As with the other criteria, this makes it unclear which measure is the most discriminating. Thus, MADRS and $QIDS\text{-}C_{16}$ differences were minimal; differences between either of these two clinician ratings and the self-report $QIDS\text{-}SR_{16}$ were slightly larger but still quite small in absolute terms.

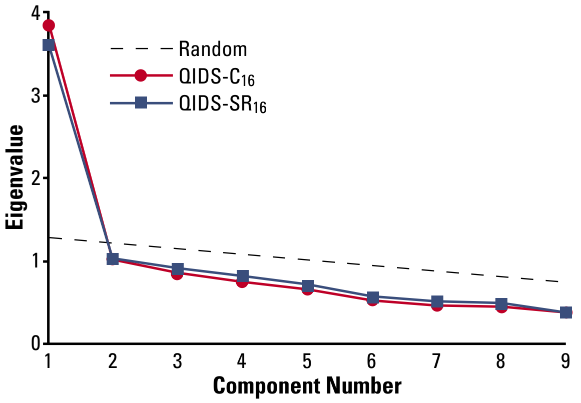
Figure 4 shows the ROC analyses. The MADRS and $QIDS\text{-}C_{16}$ were not meaningfully different. Both were slightly more accurate than the $QIDS\text{-}SR_{16}$ from a specificity of 1.0–.4 (false alarm rate of .0– .6). Past this point, the three curves converge. Table 4 shows how the three tests performed at selected thresholds.

Screening Validity Based on the MANOVA

A MANOVA using the three measures led to an F(3,274) of 55.04, *P*<.0001, so there is a multivariate difference. The more important finding is that the discriminant axis had weights of .003, .006, and .004 for the MADRS, $QIDS\text{-}C_{16}$, and $QIDS\text{-}SR_{16}$, respectively. This means that the $QIDS\text{-}C_{16}$ independently added somewhat more to overall discrimination than the other two measures, but all three did contribute separately.

FIGURE 1.

Screen plots for $QIDS\text{-}C_{16}$ and $QIDS\text{-}SR_{16}$\*



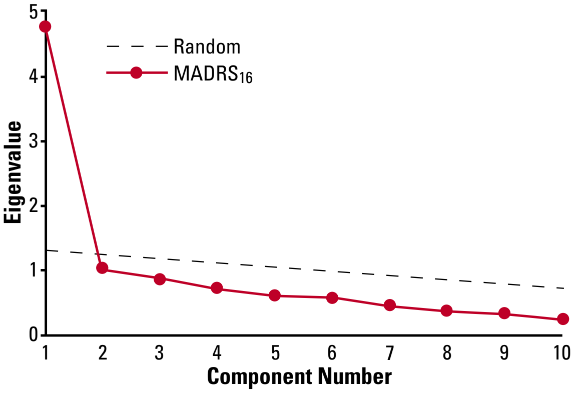
* Comparing obtained eigenvalues s with randomly generated eigenvalues (N=278).

$QIDS\text{-}C_{16}$=16-item Quick Inventory of Depressive Symptomatology–Clinician-Rated; $QIDS\text{-}SR_{16}$: 16-item Quick Inventory of Depressive Symptomatology–Self-Report.

Bernstein IH, Rush AJ, Stegman D, Macleod L, Witte B, Trivedi MH. *CNS Spectr.* Vol 15, No 7. 2010.

FIGURE 2.

Screen plots for the MADRS*



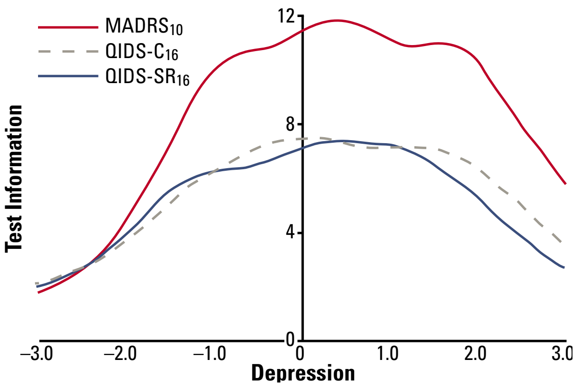
*Comparing obtained eigenvalues with randomly generated eigenvalues (N=278).

MADRS=Montgomery-Asberg Depression Rating Scale.

Bernstein IH, Rush AJ, Stegman D, Macleod L, Witte B, Trivedi MH. *CNS Spectr.* Vol 15, No 7. 2010.

FIGURE 3.

Test information for $QIDS\text{-}C_{16}$, $QIDS\text{-}SR_{16}$, and $MADRS_{10}$



N=278.

$QIDS\text{-}C_{16}$=16-item Quick Inventory of Depressive Symptomatology–Clinician-Rated; $QIDS\text{-}SR_{16}$: 16-item Quick Inventory of Depressive Symptomatology–Self-Report; MADRS=Montgomery-Asberg Depression Rating Scale.

Bernstein IH, Rush AJ, Stegman D, Macleod L, Witte B, Trivedi MH. *CNS Spectr.* Vol 15, No 7. 2010.

https://doi.org/10.1017/S1092852900000389 Published online by Cambridge University Press

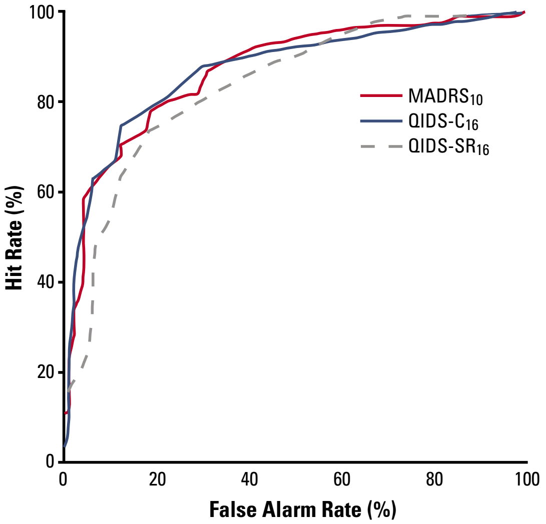
Screening Validity Based on Multivariate Logistic Regression

The MADRS contributed to overall discrimination above and beyond that afforded by the remaining two measures ($\chi^2[1]=6.8$, $P<.01$). The increments for the remaining two measures both missed achieving significance ($\chi^2[1]=3.8$ and 3.0), for the QIDS-C_{16} and QIDS-SR_{16} (.10 >P>.05.) All three measures contributed to the discrimination between depressed and not depressed patients in pairwise contrasts with associated $\chi^2(1)$ values ranging from 6.6–30.0 ($P<.05$). The increments provided by the MADRS and QIDS-C_{16} tended to be slightly larger than the increments provided by the QIDS-SR_{16}.

Test Equating

Table 5 contains the equated test scores. The test-equating process identified the observed values on each of the three scales that were closest to a common estimated depression score (theta) from the Samejima model. For example, QIDS-C_{16} and QIDS-SR_{16} scores of 12 are both equivalent to a MADRS score of 28 because all are associated with estimated depression scores of –.1, which is

FIGURE 4.
ROC curves for QIDS-C_{16}, QIDS-SR_{16}, and $MADRS_{10}$ (N=278)



QIDS-C_{16}=16-item Quick Inventory of Depressive Symptomatology–Clinician-Rated; QIDS-SR_{16}: 16-item Quick Inventory of Depressive Symptomatology–Self-Report; MADRS=Montgomery-Asberg Depression Rating Scale.

Bernstein IH, Rush AJ, Stegman D, Macleod L, Witte B, Trivedi MH. *CNS Spectr.* Vol 15, No 7. 2010.

TABLE 4.
Sensitivities, Specificities, False-Positive and Negative Rates, True-Positive and Negative Rates, and Positive and Negative Predictive Values for Selected QIDS-C_{16}, QIDS-SR_{16}, and MADRS Cutoffs (N=278)

Scale	*Cutoff*	*Sensitivity*	*Specificity*	*False Positive Rate*	*True Positive Rate*	*False Negative Rate*	*True Negative Rate*	*Positive Predictive Value*	*Negative Predictive Value*
QIDS-C_{16}	5	0.98	0.15	85	98	2	15	.68	.79
QIDS-C_{16}	7	0.95	0.33	67	95	5	33	.73	.78
QIDS-C_{16}	9	0.91	0.56	44	91	9	56	.79	.77
QIDS-C_{16}	11	0.81	0.78	22	81	19	78	.87	.68
QIDS-SR_{16}	5	0.99	0.13	87	99	1	13	.68	.87
QIDS-SR_{16}	7	0.98	0.33	67	98	2	33	.73	.89
QIDS-SR_{16}	9	0.91	0.47	53	91	9	47	.76	.74
QIDS-SR_{16}	11	0.83	0.66	34	83	17	66	.82	.67
$MADRS_{10}$	6	0.97	0.25	75	97	3	25	.71	.80
$MADRS_{10}$	8	0.97	0.33	67	97	3	33	.73	.84
$MADRS_{10}$	12	0.92	0.59	41	92	8	59	.81	.80
$MADRS_{10}$	15	0.82	0.71	29	82	18	71	.84	.68

QIDS-C_{16}=16-item Quick Inventory of Depressive Symptomatology–Clinician-rated; QIDS-SR_{16}=16-item Quick Inventory of Depressive Symptomatology–Self-report; $MADRS_{10}$=10-item Montgomery-Asberg Depression Rating Scale.

Bernstein IH, Rush AJ, Stegman D, Macleod L, Witte B, Trivedi MH. *CNS Spectr.* Vol 15, No 7. 2010.

https://doi.org/10.1017/S1092852900000389 Published online by Cambridge University Press

slightly below the sample's mean depression as inferred from the Samejima IRT model.[16,17]

DISCUSSION

Overall, the MADRS was the most reliable (L=.87) based on coefficient α and on the TIF. Both versions of the QIDS had similar and acceptable reliabilities (0.80–0.82). All three tests were unidimensional. Individual items on the two versions of the QIDS did not differ systematically. The MADRS and the QIDS-C_{16} had nearly identical (1.53 and 1.51) effect sizes in differentiating depressed from nondepressed patients, followed by the QIDS-SR_{16} (1.37). Thus, the two clinician-completed measures have a marginal psychometric advantage within the context of a univariate, linear criterion. The contribution of the QIDS-C_{16} to the MANOVA was greatest (.006) followed by the nearly equal contributions of the MADRS and

TABLE 5.
Equated QIDS-C_{16}, QIDS-SR_{16}, and MADRS Scores and Associated Estimated Depression Scores (θ)

SS	θ	SS	θ	SS	θ	SS	θ	SS	θ	SS	θ
		0	-2.4	0	-2.5	25	0.5			16	0.5
0	-2.2	1	-2.1	1	-2.2	26	0.6	16	0.6	17	0.6
1	-1.9	2	-1.9	2	-2.0	27	0.7	17	0.7		
2	-1.8	3	-1.7	3	-1.8	28	0.7				
3	-1.6			4	-1.6	29	0.8			18	0.8
4	-1.5			5	-1.4	30	0.9	18	0.9	19	0.9
5	-1.3	4	-1.4			31	1.0				
6	-1.2	5	-1.2	6	-1.2	32	1.1	19	1.1	20	1.1
7	-1.1	6	-1.1			33	1.1				
8	-1.0			7	-1.0	34	1.2	20	1.2	21	1.2
9	-0.9	7	-0.9			35	1.3				
10	-0.8			8	-0.8	36	1.4	21	1.4	22	1.4
11	-0.7	8	-0.7			37	1.5				
12	-0.6			9	-0.6	38	1.6	22	1.6	23	1.6
13	-0.5	9	-0.5			39	1.7				
14	-0.4			10	-0.4	40	1.8	23	1.8	24	1.8
15	-0.3	10	-0.4	11	-0.3	41	1.9				
16	-0.2					42	2.0	24	2.0	25	2.0
17	-0.2	11	-0.2			43	2.1				
18	-0.1	12	-0.1	12	-0.1	44	2.3	25	2.3	26	2.2
19	0.0			13	0.0	45	2.4				
20	0.1	13	0.1			46	2.5				
21	0.2			14	0.2	47	2.7	26	2.6	27	2.6
22	0.3			15	0.3	48	2.9	27	2.9		
23	0.3	14	0.3			49	3.1				
24	0.4	15	0.4			50	3.3				

Based upon N=278 patients.

$MADRS_{10}$=10-item Montgomery-Asberg Depression Rating Scale; QIDS-C_{16}=16-item Quick Inventory of Depressive Symptomatology–Clinician-rated; QIDS-SR16=16-item Quick Inventory of Depressive Symptomatology–Self-report.

Bernstein IH, Rush AJ, Stegman D, Macleod L, Witte B, Trivedi MH. *CNS Spectr.* Vol 15, No 7. 2010.

https://doi.org/10.1017/S1092852900000389 Published online by Cambridge University Press

QIDS-SR_{16} (.003 and .004). This same pattern was noted in the ROC analysis, which employs a univariate loglinear criterion.

The logistic regression weight, also a univariate loglinear criterion, for the QIDS-C_{16} was largest (.35), followed, in turn, by the QIDS-SR_{16} (.29) and MADRS (.18). This is somewhat, but not extremely, different from what was observed with the other criteria. Pairs of measures were entered in logistic regression. All three measures added to the measure with which they were paired, but the MADRS and QIDS-C_{16} produced larger increments than the QIDS-SR_{16}. Finally, the increment of each measure relative to the other two was evaluated in logistic regression. Only the MADRS significantly incremented the contribution of the QIDS-C_{16} and QIDS-SR_{16}, but the independent contributions of the latter two scales just missed achieving significance.

In a careful examination and comparison of the properties of these three scales, the MADRS tended to be best discriminate by some criteria while the QIDS-C_{16} tended to be best by other criteria. Because arguments can be made about which criteria are most relevant and because the differences were very small, the two clinician-completed measures can be viewed as not different from clinical and research perspectives. In fact, the similarities among the items suggest that results comparing the three (or other such measures based upon *DSM-IV-TR* symptoms) would probably never be extremely great, which is the case. In addition, the two clinician completed measures correlated more highly with each other than either did with the self-report QIDS-SR_{16}. This finding has been noted previously.[10,21-23] That is, clinician perspective and patient perspective are often more discrepant than two clinicians or two self-report ratings.

Study limitations include absence of data on the performance of each scale in differentiating drug from placebo in a controlled trial; the use of a single source for patients in a sample of convenience which clearly limits generalizeability; and the absence of rater blindness in most patients to the diagnosis rendered by the MINI. In fact, despite the randomization of test order, the rater would have known about the MINI results before the rating scale results in most patients. Finally, nearly all patients were on medication treatment, which may have affected the symptomatic picture.

CONCLUSION

All three tests performed well and were nearly comparable within this sample. The need to use a clinician rating should be balanced against the marginal amount of loss in validity and the time savings of self report that is nearly comparable. On the other hand, the press of daily practice and the need for a clinically useful brief measure raises the question of whether a self report (in this case the QIDS-SR_{16}) is sufficiently accurate that the time saved by a self report can be seen as an acceptable offset to the slight (in this case) loss of psychometric excellence. Looking at the sensitivities, specificities and ROC curves, as well as the overall picture presented by the findings in this study, we believe that such is the case. The MADRS and QIDS-C_{16} are basically comparable. The question becomes how far off would one be when relying solely on the self report to identify depressed cases. The answer is in Table 2. At thresholds of 7 for the QIDS-C_{16} and QIDS-SR_{16}, as well as for the MADRS at a threshold of 8, sensitivity is 97% to 98%, while specificities are 33%. At the threshold of 11 on both QIDS ratings (and 15 on the MADRS), again very similar performances are achieved. Thus, all three scales are of use both clinically and in research. Note that other brief ratings (eg, the Patient Health Questionnaire) also have established utility as a screening tool.[24] ***CNS***

REFERENCES

1. Montgomery SA, Äsberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry.* 1979;134:382-389.
2. Rush AJ, Trivedi MH, Ibrahim HM, et al. 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.* 2003;54(5):573-583. Erratum p. 585.
3. Trivedi MH, Rush AJ, Ibrahim HM, et al. 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.* 2004;34(1):73-82.
4. *Diagnostic and Statistical Manual of Mental Disorders.* 4th ed, Text Rev. Washington DC: American Psychiatric Press; 2000.
5. Sheehan DV, Lecrubier Y, Sheehan KH et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry.* 1998;59(Suppl 20):22-33.
6. First MB, Spitzer RL, Gibbon M, Williams JBW. *Structure Clinical Interview for DSM-IV Axis I Disorders - Patient Edition (SCID-I/P, version 2.0).* New York, NY: Biometrics Research Department, NY State Psychiatric Institute; 1997.
7. Svanborg P, Asberg M. A comparison between the Beck Depression Inventory (BDI) and the self-rating version of the Montgomery Asberg Depression Rating Scale (MADRS). *J Affect Disord.* 2001;64(2-3):203-216.
8. Svanborg P, Ekselius L. Self-assessment of DSM-IV criteria for major depression in psychiatric out- and inpatients. *Nord J Psychiatry.* 2003;57(4):291-296.
9. Rush AJ, Carmody TJ, Reimitz PE. The Inventory of Depressive Symptomatology (IDS): clinician (IDS-C) and self-report (IDS-SR) ratings of depressive symptoms. *Int J Methods Psychiatr Res.* 2000;9:45-59.
10. Rush AJ, Bernstein IH, Trivedi MH et al. An evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression: a Sequenced Treatment Alternatives to Relieve Depression trial report. *Biol Psychiatry.* 2006;59(6):493-501.

https://doi.org/10.1017/S1092852900000389 Published online by Cambridge University Press

11. Horn JL. An empirical comparison of various methods for estimating common factor scores. *Educ Psychol Meas*. 1965;25:313-322.
12. Humphreys LG, Ilgen D. Note on a criterion for the number of common factors. *Educ Psychol Meas*. 1969;29:571-578.
13. Humphreys LG, Montanelli RGJr. An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behav Res*. 1975;10:193-206.
14. Montanelli RGJr, Humphreys LG. Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: a Monte Carlo study. *Psychometrika*. 1976;41:341-348.
15. Lord FM. *Applications of Item Response Theory for Practical Testing Problems*. Hillsdale, NJ: LEA; 1980.
16. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychol Monogr.* 1969;4:2.
17. Samejima F. Graded response model. In: van Linden W, Hambleton RK, eds. *Handbook of Modern Item Response Theory.* New York, NY: Springer-Verlag; 1997:85-100.
18. Carmody TJ, Rush AJ, Bernstein IH, Brannan S, Husain MM, Trivedi MH. Making clinicians lives easier: Guidance on use of the QIDS self-report in place of the MADRS. *J Affect Disord.* 2006;95(1-3):115-118.
19. Carmody TJ, Rush AJ, Bernstein IH, et al. The Montgomery Asberg and the Hamilton ratings of depression: A comparison of measures. *Eur Neuropsychopharmacol.* 2006;16(8):601-611.
20. Orlando M, Sherbourne CD, Thissen D. Summed-score linking using item response theory: application to depression measurement. *Psychol Assess*. 2000;12(3):354-359.
21. Biggs MM, Shores-Wilson K, Rush AJ, et al. A comparison of alternative assessments of depressive symptom severity: a pilot study. *Psychiatry Res*. 2000;96(3):269-279.
22. Margo GM, Dewan MJ, Fisher S, Greenberg RP. Comparison of three depression rating scales. *Percept Mot Skills*. 1992;75(1):144-146.
23. Rush AJ, Giles DE, Schlesser MA, Fulton CL, Weissenburger J, Burns C. The Inventory for Depressive Symptomatology (IDS): preliminary findings. *Psychiatry Res.* 1986;18(1):65-87.
24. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. Validity of a brief depression severity masure. *J Gen Intern Med.* 2001;16:606-613.

https://doi.org/10.1017/S1092852900000389 Published online by Cambridge University Press