Item response analysis of the inventory of depressive symptomatology

Ira H Bernstein[1]
A John Rush[2]
Thomas J Carmody[2]
Ada Woo[1]
Madhukar H Trivedi[2]

[1]Department of Psychology, The University of Texas at Arlington, Arlington, TX, USA; [2]Department of Psychiatry, The University of Texas Southwestern Medical Center at Dallas, Dallas, TX, USA

Background: Both the clinician (IDS-C_{30}) and self-report (IDS-SR_{30}) versions of the 30-item Inventory of Depressive Symptomatology have acceptable psychiatric properties and have been used in various clinical studies. These two scales, however, have not been compared using item response theory (IRT) methods to determine whether the standard scoring methods are optimal.

Methods: Data were derived from 428 adult public sector outpatients with nonpsychotic major depressive disorder. The IDS-C_{30} and IDS-SR_{30} were compared using Samejima's graded response model.

Results: A model was constructed jointly fitting the IDS-C_{30} and IDS-SR_{30}. An improvement in scale performance was obtained by grouping selected items into domains (specifically sleep, psychomotor, and appetite/weight domains) analogous to the standard scoring of the 16-item Quick Inventory of Depressive Symptomatology.

Conclusions: For the IDS-C_{30} and IDS-SR_{30}, standard scoring (ie, computing total score using all individual items) provides simplicity, comparability to published data, and a basis for clinical decision making. The revised scoring method, however, improves the utility of both scales when comparing groups as it provides explicit tests of item parameters.

Keywords: Inventory of Depressive Symptomatology, item response theory, Samejima graded response model, depressive symptoms, symptom ratings

Introduction

The 30-item Inventory of Depressive Symptomatology (IDS_{30}) (Rush et al 1996, 2000; Trivedi et al 2004b) has been widely used and evaluated using classical test theory methods. The standard total score is obtained by summing the ratings of 28 of the 30 items. Either weight loss or weight gain, appetite loss or appetite gain is scored because only one member of each pair is applicable to any given respondent. Each of the 28 items is scored on a 0 to 3 scale (0–the absence of pathology; 3–severe pathology). The total scores range from 0 to 84. Standard scoring assumes a traditional model of tests known as classical test theory (CTT) in which the trait score (depression in this case) represents the scale score total plus random error of measurement. Items are the unit of analysis.

Totaling individual items is not the only way to score a test. For example, the 16-item Quick Inventory of Depressive Symptomatology ($QIDS_{16}$) (Rush et al 2000, 2003b; Trivedi et al 2004b) uses domain scoring such that when more than one item belongs to the same general domain (eg, four items assess sleep disturbance), the items are grouped and assigned a single score for that domain based upon the highest (most pathological) score for the domain-related items. Thus, for the QIDS, the scores for three domains are based on more than one item (4 items for sleep disturbance, 2 items for psychomotor disturbance, and 4 items for the appetite/weight domain). Each of the remaining 5 items is individually scored for each domain (eg, sad mood, concentration, decision making). Thus, 16 items are used to score 9 domains on the QIDS. This method allows the use of CTT analyses with the nine domains rather than items as the

Correspondence: A John Rush
Department of Psychiatry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9086, USA
Tel +1 214 648 4600
Fax +1 214 648 4612
Email john.rush@utsouthwestern.edu

units of analysis. The total score ranges from 0 to 27 rather than the 0 to 48 which would have been the case had each of 16 items been scored individually and totaled. Domain scoring avoids overcounting items in groups with a high correlation among them.

Item response theory (IRT) methods (in particular the Samejima model) (Samejima 1997) is particularly suited for graded item responses (eg, 0–3 ratings on items or domains) as with the IDS and QIDS. All IRT models scale individual items in terms of their location on an inferred continuum using a complex mathematical procedure. The underlying continuum, denoted as "Θ", refers to depression severity in this report. The unit of analysis may be items or domains.

One can employ either the CTT or IRT approach to evaluate items or domains assessed by scales like the IDS or QIDS. The more familiar CTT addresses two important aspects of scale performance. The level of response (or severity of pathology) is the item mean ($\bar{X}$). The relation of the item to overall depression is the item/total correlation (r_{it}). The larger the value of the individual item or domain, $\bar{X}$, the more severe the symptom. The higher the value of r_{it}, the more closely the rated symptom relates to overall depression.

The item (or domain) mean ($\bar{X}$) and the item (domain) total correlation (r_{it}) may not be strongly related to each other. For example, sleep disturbance items on both the IDS and QIDS generally have among the highest values of $\bar{X}$, but these sleep disturbance items are only modestly related to overall depression severity as judged by the total scale score (ie, their r_{it} values are not particularly large). Conversely, sad mood may have a lower $\bar{X}$ value, but it is more highly related to overall depression than sleep disturbance, which is expected since sad mood is a core depressive symptom (APA 2000; Bernstein et al 2006).

The IRT approach provides information not provided with CTT. IRT allows one to formally equate scores on different scales so that a total score, say X, on one depression scale can be shown equivalent to a score of Y on another. For example, we recently used IRT (Carmody et al 2006b) to equate total scores on the QIDS and the Montgomery-Äsberg Depression Rating Scale (MADRS) (Montgomery and Äsberg 1979). We did the same (Carmody et al 2006a) with the MADRS and the 17-item Hamilton Rating Scale for Depression (Hamilton 1960, 1967). Secondly, IRT allows for a comparison of groups defined, for example, by gender or other baseline demographic or clinical features in terms of both individual item responses and the frequency of different item responses in relation to overall depression severity. CTT also easily allows tests of differences in $\bar{X}$, but with CTT testing difference in r_{it} is somewhat complex (see Rush et al 2006). When specific items perform differently in different respondent groups, the term differential item functioning (dif) is used. Finally, IRT ensures a more linear relationship between the construct of depression and individual items than does the CTT approach, which might lead the resulting scores to have more optimal properties.

On the other hand, CTT methods always produce results even if the scale has undesirable properties such as low internal consistency. The IRT analyses may not be feasible in some cases because IRT analyses require stronger assumptions. For example, most IRT models assume an S-shaped relation between the magnitude of the trait and the item response. CTT analyses have been conducted with the IDS (Rush et al 1996), but IRT analyses have not been reported. This paper examined the IDS using an IRT approach.

Methods

Subjects

The sample was obtained from the Texas Medication Algorithm Project (TMAP) (Rush et al 2003a; Trivedi et al 2004a), which was conducted in accordance with international guidelines for good clinical practice and the Declaration of Helsinki. TMAP was approved by the institutional review boards at The University of Texas Southwestern Medical Center and the University of Texas, Austin, as well as by each local Institutional Review Board where applicable. All patients provided written informed consent prior to study participation.

Adult outpatients with major depressive disorder (MDD) were recruited from the public sector (Bernstein et al 2006; Trivedi et al 2004a, 2004b). The original sample of 547 outpatients with MDD was reduced to 428 by excluding those with MDD with psychotic features.

Both the self-report (IDS-SR_{30}) and clinician-rated (IDS-C_{30}) versions of the 30-item Inventory of Depressive Symptomatology (Rush et al 1996, 2000; Trivedi et al 2004b) were obtained at exit by a research coordinator not involved in patient treatment.

Statistical analysis

The goal of the analyses was to jointly fit the Samejima IRT model to the IDS-SR_{30} and the IDS-C_{30} and to evaluate differences between these two scales. We first evaluated the two scales for unidimensionality using a principal component

analysis. We compared the successive eigenvalues (scree) to those obtained by randomly generated correlations using the same number of variables and observations in a procedure known as parallel analysis (Horn 1965; Humphreys and Ilgen 1969; Humphreys and Montanelli 1975; Montanelli and Humphreys 1976). The number of components (dimensionality) is the number of components in the real data for which eigenvalues exceed those that were randomly generated.

Since each item on each scale has four response alternatives (ratings on a 0–3 scale), the Samejima model generated 4 parameters per item. One parameter describes how strongly each of 3 functions relates item (or domain) responses (ie, symptoms) to overall depression. These three functions respectively denote: (a) the tendency for a symptom to be reported as a "1", "2" or "3" relative to a "0", (b) the tendency for a symptom to be reported as a "2" or a "3" relative to a "0" or "1", and (c) the tendency for a symptom to be reported as a "3" relative to a "0", "1", or "2". The locations of the respective functions are symbolized b_0, b_1, and b_2 (collectively b_i). These locations denote the relative frequency of the dichotomized responses. A scale with mean of 0 and standard deviation 1 is common. Thus, if the estimate of b_0 to equal 0, it would imply that a "0" response is made half the time and a "1", "2", or "3" is made the remainder of the time. The slope is symbolized "a", which corresponds to the item/total correlation of CTT in measuring how strongly a given symptom domain relates to overall depression severity.

To illustrate how the IRT approach works, consider our previous work with the QIDS-SR_{16} and QIDS-C_{16}, each of which scores 9 domains (the criterion symptoms to diagnose a major depressive episode) (Bernstein et al 2006; Rush et al 2006). Both scales were found to be unidimensional. A base model was constructed pooling the two scales into a single 18-domain scale using exit data from the TMAP database (Trivedi et al 2004b). The four parameters (a, b_0, b_1, b_2) from a given item on the QIDS-SR_{16} were allowed to take on different values from the four parameters of the corresponding item on the QIDS-C_{16}, resulting in 72 free parameters (2 scales × 9 domains × 4 parameters/domain). The resulting value describes how well these 72 parameters fit the data. The individual a parameters were then tested individually by constraining each, one at a time, to be the same value in the two scales. This more constrained model also provided a goodness of fit value. The difference between the two fits is approximately distributed as a form of chi-square known as

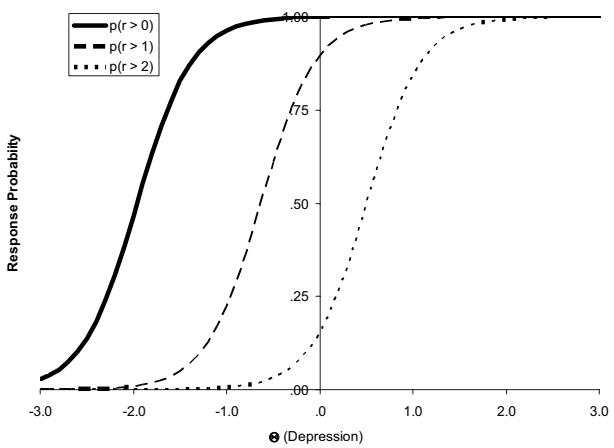

Figure 1 Boundary functions for a four-alternative item.

the likelihood-ratio chi-square (G^2) and was tested for significance with 1 df, representing the one parameter that was constrained. A significant value implies that the item slope (a) differed across the two scales. The process was repeated for each domain. Next, the a parameters were allowed to vary freely, but the three b parameters/item within each domain were constrained to equality. Values of G^2 were again obtained by comparing the value obtained from the constrained version to the value obtained from the base model. Each of these nine tests was based upon 3 df, representing the three intercepts that were constrained for each domain. A significant result implies that there are intercept differences between the two scales involving that domain. That would have meant that symptoms in that domain are reported with different frequencies by the QIDS-C_{16} and the QIDS-SR_{16}. In fact, no significant slope differences were found, and only one intercept difference was found (for agitation/retardation). When slopes or intercept differences are found between groups or measurement methods (in this case), the term differential item functioning (DIF) is used. In the case of the QIDS, only the agitation/retardation domain performed differently when one scale as opposed to another was used. Even then, the difference was not in the degree of relationship between the symptom and overall depression (ie, the a parameter was not different). Rather, patients self-reported slightly greater psychomotor disturbance than did clinical raters.

All IRT analyses with the $QIDS_{16}$ have been consistent. No study has produced an anomalous result. An anomalous result means that a better fit is found with a more constrained than a less constrained model, which leads to a spurious "negative" G^2. Such a result can arise from various sources:

(a) very high correlations between individual items, (b) small cell frequencies, or (c) long scales.

When we conducted similar analyses with the 30-item IDS scales, such anomalies were encountered. Consequently, we made the following modifications to successfully fit the model and to test for DIF: (a) we replaced scoring of individual items with domain scores for sleep, psychomotor, and appetite/weight domains; (b) we pooled items with few positive responses; and (c) we changed how the base model was tested.

Specifically, IDS items 1–4 were combined into a single sleep domain; items 11–14 were combined into a single appetite/weight domain, and items 23–24 were combined into a single psychomotor domain (analogous to the standard scoring of the $QIDS_{16}$) (Rush et al 2003b). This scoring resulted in 23 domains.

Next, items 6–8, 16, and 21 (diurnal mood variation, distinct quality to mood, distinct mood quality, interest in sex, and gastrointestinal complaints) were dichotomized into 0 vs 1 or greater because responses of 2 or greater to each of these items were rare. Finally, the tested strategy was reversed by first generating a base model in which all parameters were constrained to equality and then freeing a parameters individually and b_i parameters in groups of 3. This procedure (the converse of what was used with the $QIDS_{16}$) maintains the idea of comparing more vs. less constrained models. As an addendum to this testing, individual b_i parameters were tested specifically by freeing them whenever the entire group of three parameters differed. For example, if there was a difference in the overall distributions of the sad mood response frequencies between the two rating scales, more specific differences involving the three specific dichotomies (0 vs 1 to 3, 0 or 1 vs 2 or 3, and 0 to 2 vs 3) were examined individually.

We next conducted CTT analyses using the 23 domains and all 28 items. Finally, the test information functions of the IDS and the QIDS (obtained by extracting the relevant items from the IDS were compared. In the present context, the test information function (TIF) describes how well a test can discriminate small differences in depression as a function of the score— the higher the value, the more discriminating the test. TIF bears similarities to the internal consistency (coefficient alpha) obtained by CTT, but the TIF reveals how test information varies over different levels of depression rather than being computed as a constant.

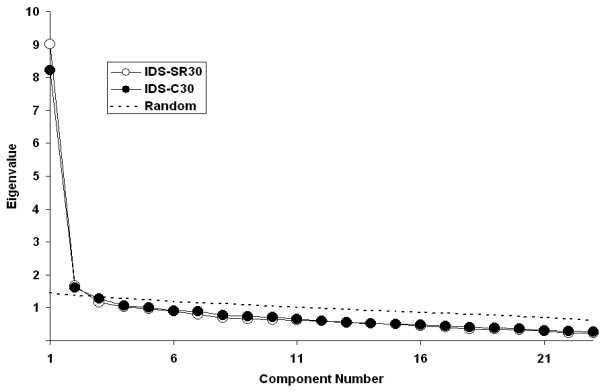

Figure 2 Plot of successive eigenvalues (scree) for the IDS-C_{30}, IDS-SR_{30}, and randomly generated data (parallel analysis).

Results

Dimensionality

Figure 2 contains the successive eigenvalues of the IDS-C_{30} and IDS-SR_{30} (scree) and those randomly generated. Note that both scales meet the criteria for two factors since the first two eigenvalues exceed the randomly generated data. The disparity between the real and randomly generated second eigenvalue, however, was modest. The presence of two factors imposes a limitation on the IRT solution to be provided, since that solution assumes unidimensionality. To the extent that depression is represented by the first principal component, the items related to the second principal component define something other than depression that is contributing to the score.

Table 1 contains the first and second principal component loadings for the IDS-SR_{30} and IDS-C_{30}, respectively. Using an arbitrary cutoff of 0.4 to denote a large weight on a given component, the second principal component of the IDS-SR_{30} is defined by aches and pains (domain 18), symptoms of sympathetic nervous system arousal (domain 19), and gastrointestinal complaints (domain 21). These same three items plus the presence of the capacity for pleasure (negative loading) (domain 15) formed the second principal component for the IDS-C_{30}. Thus, the three items common to both forms deal with somatic symptoms that may not necessarily reflect depression per se, especially in this sample of socially disadvantaged individuals with high rates of general medical conditions (Trivedi et al 2004a).

Scores on the first and second components were then generated for the IDS-SR_{30} and IDS-C_{30}. As expected, the correlation between the first component scores for the IDS-SR_{30} and IDS-C_{30} was extremely high ($r = 0.92$) because both

Table 1 Principal component structure, variance accounted for (h^2), factor variances for the IDS-SR_{30} and the IDS-C_{30} (23 domains scored)

	IDS-SR_{30}			IDS-C_{30}		
Domain	**I**	**II**	**h^2**	**I**	**II**	**h^2**
Sleep	0.50	0.11	0.26	0.54	0.18	0.32
Sad mood	0.82	–0.09	0.68	0.82	–0.11	0.68
Irritability	0.68	0.08	0.47	0.62	0.06	0.39
Anxiety/Tension	0.75	0.17	0.60	0.70	0.21	0.54
Mood reactivity	0.68	–0.32	0.56	0.70	–0.31	0.59
Diurnal variation	0.24	0.13	0.08	0.31	0.32	0.20
Distinct quality to mood	0.51	–0.27	0.33	0.30	–0.17	0.12
Appetite/Weight	0.45	–0.03	0.20	0.41	0.02	0.17
Concentration/Decision making	0.72	0.01	0.53	0.64	–0.03	0.41
Self view	0.69	–0.22	0.53	0.71	–0.17	0.53
Future view	0.72	–0.30	0.60	0.72	–0.28	0.60
Thoughts of death and suicide	0.57	–0.11	0.34	0.61	–0.19	0.40
General interest	0.74	–0.30	0.64	0.76	–0.27	0.65
Energy level	0.74	–0.11	0.56	0.72	0.00	0.52
Capacity for pleasure[a]	0.76	–0.35	0.70	0.68	–0.41	0.63
Interest in sex	0.52	–0.30	0.36	0.54	–0.20	0.33
Restlessness/Agitation	0.63	0.20	0.43	0.58	0.21	0.38
Somatic complaints[b]	0.53	0.42	0.46	0.46	0.51	0.47
Sympathetic arousal[b]	0.59	0.56	0.66	0.57	0.48	0.56
Panic/Phobia	0.59	0.36	0.47	0.54	0.17	0.32
Gastrointestinal complaints[b]	0.35	0.43	0.31	0.32	0.46	0.31
Interpersonal sensitivity	0.62	0.16	0.41	0.54	–0.07	0.30
Leaden paralysis	0.64	0.29	0.49	0.57	0.31	0.42
Factor variance	0.39	0.07	0.46	0.36	0.07	0.43

[a]The presence of the capacity for pleasure also contributes to the second component for only the IDS-C_{30}.
[b]These three items contribute to the second principal component for both the IDS-C_{30} and the IDS-SR_{30}.
Note: Since the results were obtained from the principal components, the variances accounted for in each item (h^2) are the sum of squared structure elements, eg, $0.50^2 + 0.11^2 = 0.26$ for domain 1 on the IDS-C_{30}.

represent the dominant depression component of both scales. The correlation between the two second component scores was also moderately high (r = 0.73). The remaining correlations (eg, between the first component of the IDS-SR_{30} and the second component of the IDS-C_{30}) were 0.11 or less, which establishes the independence between the first and second components of each scale. These findings indicate that the multidimensionality within the IDS-C_{30} and IDS-SR_{30} is consistent.

IRT model parameters

Table 2 shows the Samejima a and b_i parameter estimates for the IDS-SR_{30} and IDS-C_{30} when they were each scored to create 23 domains. The last column (Diff.) identifies domains for which there is a significant difference between the clinician and self-report ratings in the a or b_i parameter estimates (ie, differential item functioning) (DIF). Note that three of these are also QIDS items (concentration/decision making, capacity for pleasure, restlessness/agitation) vs one that is peculiar to the IDS (diurnal variation). None of the former provided any evidence of a slope difference when scored as part of the QIDS. Therefore, it is not the item itself but rather the broader definition of depression used by the IDS that accounts for these differences.

Scoring via CTT and IRT

Each version of the 23 domain-scored IDS may be scored two ways – by CTT simply as the as the sum of the 23 domains or by a fairly complex IRT algorithm. The correlation between the CTT and IRT scores on each of the two scales is high (0.92 for CTT and 0.91 for IRT). The correlation between the two methods of scoring is even higher, 0.97, for both versions of the scale. However, the CTT and IRT methods of scoring the 23 domain-scored versions of the IDS-SR_{30} and IDS-C_{30} reveal that total scores using the two scoring methods are not linearly related.

Table 2 Item response theory (IRT) parameter estimates for the 23 domain versions of the IDS (IDS-C_{28} and the IDS-SR_{28})

	IDS-C_{30}				IDS-SR_{30}				
Domain	**a**	**b_0**	**b_1**	**b_2**	**a**	**b_0**	**b_1**	**b_2**	**Diff.**
1. Sleep	1.17	–2.62	–1.48	0.00	1.17	–2.62	–1.48	0.00	
2. Sad mood	2.85	–0.98	0.11	1.03	2.85	–0.98	0.11	1.03	
3. Irritability	1.59	–0.77	0.58	1.83	1.59	–0.77	0.58	1.83	
4. Anxiety/Tension	1.85	–1.20	0.00	1.18	1.85	–1.20	0.00	1.18	
5. Response to events	1.99	0.10	0.73	1.43	1.99	0.10	0.56	1.43	b_2
6. Diurnal variation	0.55	0.78			0.55	0.78			
7. Distinct quality to mood	0.55	–0.52			1.65	–0.52			a
8. Appetite/Weight	0.74	–1.24			0.74	–1.24			
9. Concentration/Decision making	1.58	–0.57	0.38	1.40	1.96	–0.57	0.38	1.40	a
10. Self view	2.07	0.02	0.72	1.29	2.07	0.02	0.72	1.09	b_3
11. Future view	1.97	–0.28	0.58	1.36	1.97	–0.54	0.58	1.36	b_1
12. Thoughts of death/Suicide	1.80	0.53	1.69	2.90	1.80	0.53	1.69	2.90	
13. General interest	2.23	–0.24	0.50	1.22	2.23	–0.24	0.50	1.22	
14. Energy level	1.95	–0.52	0.32	1.24	1.95	–0.52	0.32	1.24	
15. Capacity for pleasure	1.98	0.18	0.70	1.25	2.71	–0.13	0.70	1.43	a, b_3
16. Interest in sex	1.12	–0.36			1.12	–0.36			
17. Restlessness/Agitation	1.54	–0.94	0.67	3.30	1.10	–0.94	0.67	1.84	a, b_3
18. Aches and pains	0.92	–1.89	0.01	1.26	0.92	–1.89	0.01	1.26	
19. Sympathetic arousal	1.17	–1.01	0.81	2.05	1.17	–1.01	0.81	2.05	
20. Panic/Phobia	1.23	0.27	0.88	2.25	1.23	–0.07	1.16	1.78	b_1,b_2,b_3
21. Gastrointestinal complaints	0.66	0.59			0.66	0.59			
22. Interpersonal sensitivity	1.34	0.07	0.61	1.66	1.34	–0.30	0.90	1.66	b_1,b_2
23. Physical energy	1.42	0.06	0.89	1.70	1.42	0.06	0.89	1.70	

Notes: Diff = parameter estimates that differ between the clinical and self-report versions, ie, exhibit dif.

Figure 3 shows a scatter plot of the two sets of IDS-C_{30} scores, with IRT-generated scores along the abscissa and CTT-generated score along the ordinates. According to IRT, the IRT-generated scores are, by definition, linearly related to depression (Θ). Consequently, the CTT-generated scores are an ogival (S-shaped) function of depression. If one accepts the IRT scoring of "true" depression, then scores at the high and low ends of the CTT scale represent less difference in depression than scores in the middle. Scores in the middle of both scales are linearly related to one another. In other words, very low scores as defined by CTT tend to somewhat underestimate depression, as seen in the minimal changes in such scores following changes in IRT-generated scores. Similarly, very high scores defined by CTT tend to overestimate depression. However, this effect is modest and may or may not be of clinical significance.

The next step was to compare the results of standard 28-item CTT scoring to the two alternative 23 domain scoring methods (CTT and IRT). The 28-item scoring correlated greater than 0.999+ with the 23-item CTT domain scoring. Thus, the two CTT methods correlated to nearly identical degrees with the IRT scoring. Of course, correlations between the 28-item CTT scoring with IRT scoring were the same as the correlations between the 23-domain CTT scoring and IRT scoring (0.97). Thus, despite the curvilinearity of CTT scoring, the major properties of both CTT and IRT scoring, such as the rank-ordering of individuals, were preserved.

Finally, test information functions (TIFs) were generated for the IDS and the QIDS for the 23 domain scored versions

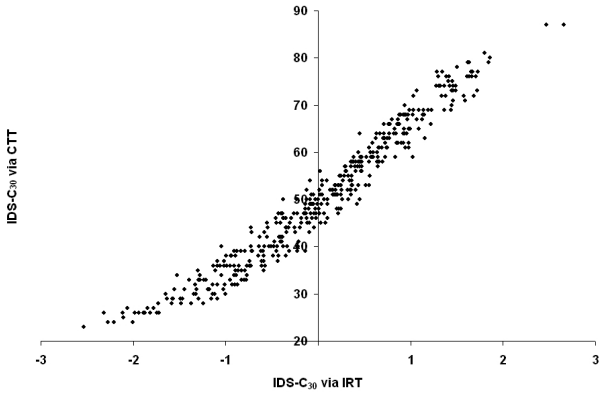

Figure 3 Scatter plot of the IDS-C_{30} scored by item response theory (IRT) vs the IDS-C_{30} scored by classical test theory (CTT).

of both the self-report and clinical IRT versions. These functions describe to what degree change in the level of depression is reflected in changes in IRT-defined test scores. The formula for the test information function may be found in Nunnally and Bernstein (Nunnally and Bernstein 1994, p 408) and Lord (Lord 1980, p 68). The TIF serves a role that is similar to the internal consistency (coefficient alpha) of CTT, but it is a function of the trait being investigated rather than a constant for the test as a whole.

Figure 4 contains the resulting functions for the clinician versions of the test. The self-report versions gave highly similar results. Thus, the IDS-C_{30} and IDS-SR_{30} relate similarly to overall depression when scored using the 23-domain method.

As can be seen, the test information of the IDS is approximately twice that of the QIDS at each level of depression (generically symbolized "Θ" in the language of IRT). This means that the IDS provides a more sensitive measure of depression than the QIDS. Technically, this applies to the IRT-scored (23 domain) version of the IDS, but the similarities between this scoring and the 28 item scoring plus the clinical and self-report versions make this a rather general conclusion. This result is expected given the greater length and greater breadth of symptom coverage with the IDS. Also, note that both tests are maximally sensitive with patients who are of average depression to one standard deviation above average in this sample, which means that it is less useful in discriminating among patients low in depression (remitted depressives and normals) and those who are extremely depressed. These results may reflect, in part, that relatively few depressives were remitted in this sample.

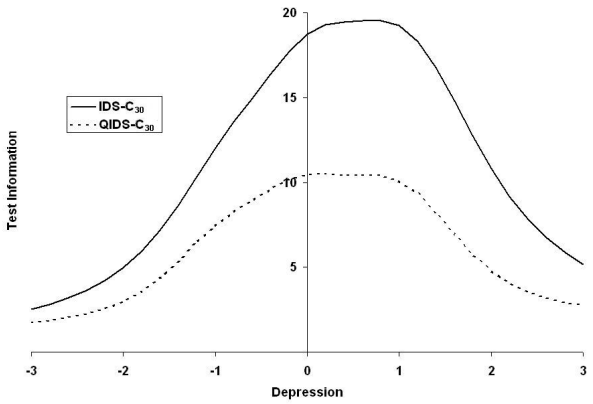

Figure 4 Test information functions for the IDS-C_{30} and the QIDS-C_{30}.

Discussion

The Samejima model was applied to the clinician and self-rated versions of the IDS. As noted above, this model greatly facilitates statistical comparisons. Whereas classical methods allow the frequency of symptoms to be compared with ease, classical methods are less suitable to evaluate differences in the relation of symptoms to depression (or any other trait). It is equally easy to evaluate both types of relationships using IRT. However, with the IDS there was greater difficulty and more strain on the assumption of unidimensionality than was the case with the QIDS, which, as our earlier papers have shown, was clearly unidimensional (Bernstein et al 2006; Rush et al 2006).

These data suggest that scoring the IDS conventionally is satisfactory for making judgments about patient care. Conventional scoring (ie, totaling the 28 items) allows one to take advantage of the far simpler CTT scoring. It would appear that creating 23 domains offers little advantage because of the relatively large number of items. On the other hand, an IRT model generated using the 23-domain scoring method, is important for virtually any research involving the IDS, since it means that groups and/or conditions can be compared using the relatively straightforward methods based upon testing for fit differences. This is a major advantage in looking for differences in the relation of domains to depression as a whole. For example, when comparing patients with postpartum depression and depression outside the postpartum period, the revised IDS scoring (23 domains) is preferred. At the same time, the fact that CTT and IRT scoring lead one to comparable results means that conventional scoring of the IDS is appropriate.

It is reasonable to ask if the revised (23 domain) scoring in either CTT or IRT form is sufficient. The answer is that it is. At the same time, this revised scoring does not seem to be more sensitive to mild depression.

One could also ask about jettisoning those items that seem to induce multidimensionality (ie, gastrointestinal symptoms, somatic complaints, sympathetic arousal). Theoretically, this makes sense. However, the loss in comparability with previous studies would probably be greater than any largely theoretical gain. Finally, there is the question of whether it is good for many of the items that load on the first principal component to reflect anxiety. That question seems very difficult to answer without considering whether anxiety is or is not an inherent part of depression or whether a large sample of depressed patients could have included some with "anxious depression" and others with minimally or non-anxious depression.

Limitations

Limitations in the assertions that an IRT model can be created for the IDS_{30} and that it is consistent across methods needs be noted. First, there is a slight, but consistent, degree of multidimensionality in the IDS_{30} that is counter to the assumption of unidimensionality made in standard usage of the Samejima model. The precise nature of the second dimension might be different in a sample with less medical comorbidity. Clearly, a replication is called for with additional patient samples.

Conclusions

Standard scoring of the IDS-C_{30} and IDS-SR_{30} provides simplicity and comparability to published data. The present results based on the IRT model enhance the validity of comparisons between groups or conditions. While clinical decisions can be made about patients with standard scoring (ie, totaling all items to obtain a scale score), researchers may wish to use this revised IRT scoring method to improve the use of the IDS when comparing groups.

Acknowledgments

This research was supported by NIMH Collaborative Grant "Development of the Inventory of Depressive Symptomatology," (MH-68851) to UT-Southwestern Medical School (A John Rush PI), UT-Arlington (Ira H. Bernstein PI), and Duke University (P Murali Doraiswamy PI).

References

[APA] American Psychiatric Association 2000. Diagnostic and Statistical Manual of Mental Disorders, 4th ed, text revision. Washington DC: APA.

Bernstein IH, Rush AJ, Carmody TJ, et al. 2006. Clinical vs. self-report versions of the Quick Inventory of Depressive Symptomatology in a public sector sample. *J Psychiatr Res*, doi:10.1016/j.jpsychires.2006,04.001.

Carmody TJ, Rush AJ, Bernstein IB, et al. 2006a. The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. *Eur Neuropsychopharmacol*, 14:601–11.

Carmody TJ, Rush AJ, Bernstein IH, et al. 2006b. Making clinicians lives easier: guidance on use of the QIDS self-report in place of the MADRS. *J Affect Disord*, 95:115–18.

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*, 6:278–96.

Horn JL. 1965. An empirical comparison of various methods for estimating common factor scores. *Educ Psychol Meas*, 25:313–22.

Humphreys LG, Ilgen D. 1969. Note on a criterion for the number of common factors. *Educ Psychol Meas*, 29:571–8.

Humphreys LG, Montanelli RG Jr. 1975. An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behav Res*, 10:193–206.

Lord FM. 1980. Applications of item response theory for practical testing problems. Hillsdale, NJ: LEA,.

Montanelli RG Jr, Humphreys LG. 1976. Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: a Monte Carlo study. *Psychometrika*, 41:341–8.

Montgomery SA, Äsberg M. 1979. A new depression scale designed to be sensitive to change. *Br J Psychiatry*, 134:382–9.

Nunnally JC, Bernstein IH, 1994. Psychometric theory. New York: McGraw-Hill.

Rush AJ, Bernstein IH, Trivedi MH, et al. 2006. An evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression: a Sequenced Treatment Alternatives to Relieve Depression trial report. *Biol Psychiatry*, 59:493–501.

Rush AJ, Carmody TJ, Reimitz PE. 2000. The Inventory of Depressive Symptomatology (IDS): clinician (IDS-C) and self-report (IDS-SR) ratings of depressive symptoms. *Int J Methods Psychiatr Res*, 9:45–59.

Rush AJ, Crismon ML, Kashner TM, et al. 2003a. Texas Medication Algorithm Project, phase 3 (TMAP-3): rationale and study design. *J Clin Psychiatry*, 64:357–69.

Rush AJ, Gullion CM, Basco MR, et al. 1996. The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychol Med*, 26:477–86.

Rush AJ, Trivedi MH, Ibrahim HM, 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*, 54:573–83. Erratum p 585.

Samejima F. 1997. Graded response model. In van Linden W, Hambleton RK (eds.). Handbook of modern item response theory. New York: Springer-Verlag. p 85–100.

Trivedi MH, Rush AJ, Crismon ML, et al. 2004a. Clinical results for patients with major depressive disorder in the Texas Medication Algorithm Project. *Arch Gen Psychiatry*, 61:669–80.

Trivedi MH, Rush AJ, Ibrahim HM, 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.