Reliability and Validity of the Turkish Version of the Hamilton Depression Rating Scale

A. Akdemir, M.H. Türkçapar, S.D. Örsel, N. Demirergi, I. Dag, and M.H. Özbay

The aim of the study was to examine the reliability and validity of the Turkish version of the Hamilton Depression Rating Scale (HDRS). Ninety-four patients with major depression/depressive mood disorders and 40 healthy controls participated in the study. The severity of depression was assessed with the HDRS, Beck Depression Inventory (BDI), and Clinical Global Impression score (CGI). The test-retest reliability coefficient of the HDRS was based on a 5-day interval was .85, with a Cronbach alpha coefficient of .75 and a split-half reliability coefficient of .76. Interrater reliability coefficients based on the independent ratings of four assessors were between .87 and .98. The correlation between the HDRS and BDI scores was .48, and between the HDRS and CGI it was .56. Principal Components Analysis yielded six factors. The correlation (−.13) between the control and patient groups indicates that the HDRS assesses depression very well.

THE HAMILTON Depression Rating Scale (HDRS) is used to measure the severity of depressive symptoms.[1] Although this scale has been used in many studies, its validity and reliability has been questioned and criticized.[2,3] The HDRS is a standard scale based on psychiatrists' assessments, and was developed in the late 1950s to measure depressive symptoms.[1] The scale was initially designed to obtain a total score based on 17 of its 21 items. The 17-item version of the scale was modified by Max Hamilton. It has been used widely in research for initial and follow-up assessments of depressive symptoms.[4] It also has practical value in the evaluation of the results of therapy. The scale was initially composed of open-ended questions directed to the patient by the evaluator. Afterward, the scale was modified to include standard questions for each item.[4]

This study examines the validity and reliability of the HDRS in the Turkish population, which has been widely used in clinical depression studies. The characteristics of our population that had effects on the results of this study are also discussed in particular.

METHOD

Study Sample

The study was performed in both the outpatient and inpatient services of the SSK Ankara Residency Training Hospital Psychiatry Clinic between September 1994 and January 1996. The sample consisted of 94 consecutive patients with any DSM-III-R[5] depressive mood disorder (major depression, dysthymia, and depressive period of bipolar depression and major depression and dysthymia) diagnosed by the Structured Clinical Interview for DSM-III-R (SCID).[6,7] Patients with a comorbid axis I diagnosis such as alcohol abuse/dependency, substance abuse/dependency, schizophrenia, delusional disorder, or eating disorder, and patients who were receiving any psychopharmacologic treatment in the prior month, were excluded from the study. Informed consent was obtained from all patients. Sixty-two of the subjects (66%) were female and 32 (34%) were male. The mean age was 32.8 ± 11.7 years (SD) for females, 39 ± 9.5 for males, and 34.9 ± 10.6 for the total group. The control group included 44 healthy individuals (13 males and 31 females) who were matched by age, sex, and socioeconomic status. Their mean age was 37.4 years. Other demographic features of the sample are presented in Table 1.

Instruments

All patients were administered the Turkish translation of the HDRS. The HDRS is an interview rating scale consisting of 17 items. The scale is used by determining the presence or absence of the symptoms in each item and ranking them as mild, moderate, or severe via the psychiatrist's questions for the related item and assessing the answers. The total score of the scale is 0 to 53, obtained by summing the ratings. A structured interview guide for the HDRS (SIGH-D) was used for interviewing patients.[4] The Clinical Global Impression scale (CGI) was used to assess the severity of depression.[8] The Beck Depression Inventory (BDI) was used for 38 patients. It is a valid self-rating depression scale in the Turkish population.[9,10] The SIGH-D was translated into Turkish and the Turkish translation was retranslated into English by two independent translators. In the trial of this reorganized scale, administration of the scale was observed through a one-way mirror by four psychiatrists who will later be included in the inter-rater reliability test, and then the corrections on the scale were completed.

Procedure

All patients completed a demographic questionnaire developed by the investigators. Then, they were interviewed using the Turkish version of the SCID for axis I diagnosis. The SCID was also used to assess depression severity and the severity of

From the Department of Psychiatry, SSK Ankara Training Hospital, Ankara; and Faculty of Psychology, Hacettepe University, Ankara, Turkey.

Address reprint requests to M.H. Türkçapar, M.D., Menevis S. 83/15 A.Ayranci, 06690 Ankara, Turkey.

0010-440X/01/4202-0014$35.00/0

doi:10.1053/comp.2001.19756

Table 1. Sociodemographic and Clinical Features of the Sample

	Depressed Patients (n = 94)		Control Group (n = 40)	
	No.	%	No.	%
Educational level				
Literate	11	11.7	2	4.9
Primary school graduate	36	38.3	18	43.9
Secondary school graduate	40	42.6	14	36.7
University graduate	7	7.4	6	14.5
Marital status				
Married	60	63.8	37	84.1
Single	34	36.2	7	15.9

depressive disorders (0 = none, 1 = mild, 2 = moderate, 3 = severe, 4 = severe with psychotic features [mood congruent], and 5 = severe with psychotic features [mood incongruent]). In our study, since patients in remission are not included in the sample, the severity of depression is determined by rating on five points. Then, the data collection methods of the study were applied to the patients. During administration of the HDRS, 40% of patients agreed to be videotaped.

Statistical Analysis

The data were analyzed by the SPSS for Windows 5.01 statistical package (SPSS, Chicago, IL). Independent sample *t* test was used for comparison of two groups, and one-sided analysis of variance was used in multiple-group comparisons of quantitative variables, and qualitative variables tested by the chi-square test. Internal consistency was calculated with the Cronbach alpha test. The Pearson correlation test was used in the calculation of the reliability coefficients and validity of similar scales, and principal-components factor analysis was performed for the investigation of structural validity. The significance level (*P* value) for statistical procedures was .05.

RESULTS

Demographic data and comparisons of subdiagnoses are presented. Also, the results of validity and reliability tests are summarized.

General Results

The mean ± SD and range for the total scores obtained from the scales are as follows: HDRS and CGI mean scores were 21.8 (SD 6.89) and 4.36 (SD 1.16) respectively. BDI mean score was 30.0 (SD 11.43). Ninety patients (95.7%) were diagnosed with major depressive disorder by the SCID. Other DSM-III-R diagnoses were dysthymia (n = 3) and bipolar disorder depressive type (n = 1). Twelve of 94 patients had another comorbid axis I diagnosis. Fifty-four patients had only one depressive episode, whereas 26 patients had two and 14 patients had three episodes. The mean number of episodes was 1.69 in males and 1.82 in females. The number of episodes was not related to the severity of depression. Although the educational level, marital status (Table 1), and presence or absence of comorbid diagnosis did not exert a significant effect on HDRS and BDI scores, both were shown to be affected by sex variable (Table 2).

Comparison of the Control Group and Patient Group

There was no significant difference between the control and patients groups with regard to the age ($t = -1.30$, $P = .20$), sex ($P = .56$), education ($t = -1.16$, $P = .25$), marital status ($P = .11$) and income ($P = .56$). The correlation for the HDRS between two groups was −.13.

Reliability Investigations

The Cronbach alpha internal consistency coefficient of the 17-item HDRS was .75 and the split-half reliability coefficient was .76 by the Spearman Brown formula.

Test-Retest Reliability

The HDRS was administered again to 93 patient 5 days after the first assessment. The raters at the retest assessments were blind to the initial ratings. One patient was kept out of this analysis because

Table 2. Means and Standard Deviations of HDRS Items and Total Scores in Respect to Sex and *t* Test Comparisons Between the Sexes

HDRS Item	Male (mean ± SD)	Female (mean ± SD)	*t*	*P*
Depressed mood	2.06 ± 0.76	2.24 ± 0.82		
Work and activities	2.47 ± 1.24	2.47 ± 1.19		
Genital symptoms	1.38 ± 0.83	1.26 ± 0.79		
Somatic, gastrointestinal	1.03 ± 0.82	1.16 ± 0.75		
Insomnia, early	1.25 ± 0.95	1.47 ± 0.80		
Insomnia, middle	0.78 ± 0.87	1.29 ± 0.84	−2.76	.0007
Insomnia, late	0.69 ± 0.86	0.98 ± 0.98		
Somatic general	1.41 ± 0.80	1.63 ± 0.63		
Feelings of guilt	1.09 ± 1.03	1.34 ± 0.94		
Suicide	1.16 ± 1.17	1.66 ± 1.21	−1.94	.056
Anxiety, psychic	1.59 ± 0.98	2.00 ± 0.96	−2.76	.0007
Anxiety, somatic	1.41 ± 1.04	1.79 ± 0.91		
Hypochondriasis	1.13 ± 0.98	0.84 ± 0.91		
Insight	0.63 ± 0.55	0.65 ± 0.52		
Retardation	1.38 ± 0.71	1.26 ± 0.79		
Agitation	0.97 ± 0.70	1.13 ± 0.80		
Loss of weight	0.88 ± 0.88	1.16 ± 0.89		
HDRS total score	21.28 ± 8.22	24.16 ± 5.93	−1.76	.054

he did not show up for the control visit. The test-retest reliability coefficient was calculated as $r = .85$, the correlation between the total scores obtained from the two administrations. Test-retest correlations results are summarized in Table 3.

Interrater Reliability

Four psychiatrists watched the video recordings of 40 patients and rated the HDRS independently. Correlations of the total scores based on the psychiatrists' ratings were found to be high; the interrater reliability coefficient ranged between .87 and .98 and all correlations were highly significant ($P < .0001$; SD 39).

Validity Investigations

Validity of similar scales. Correlations between the BDI, CGI, and HDRS were calculated to establish a basis for concurrent validity, which is generally mentioned as "validity of similar scales." The HDRS-BDI total score correlation was .48 ($P < .005$); the HDRS-CGI total score correlation was .56 ($P < .0001$). In addition, the correlation between the HDRS total score and rating score of 1 to 5 obtained by the SCID, used in assessing the severity of depression, was found to be .37 ($P < .0001$).

When the sample was divided into three subgroups with regard to CGI scores based on clinicians' observations as mild (CGI score 1 to 3), moderate (CGI score 4), and severe depression (CGI score 5 to 7), the HDRS-BDI total score correlation was found to be statistically significant only in the moderately depressed group ($r = .66$, $P < .01$).

Structural validity. Principal components factor analysis was applied to the data obtained from 94 subjects to investigate the structural validity of the HDRS. At the end of this analysis, real values of six factors were greater than 1 in respect to Kaiser normalization criteria. After orthogonal varimax rotation, factor patterns, real values, and percent variances of these factors explain 61.3% of the total variance (Table 4).

Table 3. Test-Retest Correlations of HDRS Items

Item	*r*	Item	*r*
1. Depressed mood	.61	10. Suicide	.67
2. Work and activities	.73	11. Anxiety, psychic	.80
3. Genital symptoms	.76	12. Anxiety, somatic	.79
4. Gastrointestinal	.71	13. Hypocondriasis	.79
5. Insomnia, early	.69	14. Insight	.79
6. Insomnia, middle	.79	15. Retardation	.85
7. Insomnia, late	.76	16. Agitation	.66
8. Somatic general	.66	17. Loss of weight	.08
9. Feelings of guilt	.78	Total score	.85

Table 4. Patterns, Real Values (E), and Percent Variances of the Factors Obtained by the Application of Principal Components Analysis to the HDRS

Factor	Item	Item
1. Agitating depression	16. Agitation	.79
	4. Somatic gastrointestinal (appetite)	.71
E = 3.68	17. Loss of weight	.54
V = 21.6%	1. Depressed mood	.51
	11. Anxiety, psychic	.48
2. Anxious depression	12. Anxiety, somatic	.75
	10. Suicide	.71
E = 188	9. Guilt	.70
V = 11.1%	11. Anxiety, psychic	.45
3. Vegetative depression	6. Insomnia, middle	.83
	5. Insomnia, early	.65
E = 1.39	7. Insomnia, late	.55
V = 8.2%	17. Loss of weight	.38
4. Somatized depression	13. Hypochondriasis	.66
	3. Genital symptoms	.65
E = 1.39	1. Depressed mood	.36
V = 7.2%	7. Insomnia, late	.36
5. Retardation	15. Retardation	.90
	2. Work and interests	.54
E = 1.16	7. Insomnia, late	.33
V = 6.8%	1. Depressed mood	.31
6. Miscellaneous	14. Insight	.71
	8. General somatic symptoms	.70
E = 1.07		
V = 6.3%	7. Insomnia, late	.31

DISCUSSION

Since the HDRS is a nonstructured scale, there are some criticisms that it is not valid and reliable.[11] In an attempt to standardize the HDRS, Williams developed a structured form called the SIGH-D. We adapted this scale into Turkish and used this version in our study.

In this study, 66% of the patients were females and their mean total scores of the HDRS and BDI were higher than that of males. Table 2 shows that females had higher scores on "psychic anxiety" and middle insomnia. This difference is statistically significant and the level of significance is close to "suicide" item and "total score." In another Turkish study, no sex difference was found in patients with the diagnosis of somatization disorder and depressive disorders.[12] The difference in

the results of these two studies may arise from sampling and evaluation differences. In this study, it was shown that the HDRS scores were not affected by demographic factors, such as level of education, marital status, and monthly income.

The interval between test-retest administrations was cut down to 5 days to decrease the amount of days without treatment. We obtained a correlation coefficient of .85 between the scores of the two administrations. This showed a totally consistent rating in time. When we look at the each item of HDRS one by one in retest evaluation (Table 3), the weight loss item naturally showed the weakest correlation (.08) compared with the correlations of other items (.61 to .85). This low correlation resulted from the changing of the score of this item in the retest evaluation.

The inter-rater reliability coefficients of the scale were found to be high (.87 to .98). Cicchetti and Prusoff investigated inter-rater reliability of HDRS,[13] inter-rater correlation increased from 0.46 to 0.82 upon administering in the scale in individual interviews. In another study, the concordance between the two psychiatrists was found to be 0.89.[14]

Internal consistency and split-half reliability analysis showed strong consistency of the items of the scale with each other and gave satisfactory results. The reliability coefficients of the scale were between .75 and .76. Thus, the items of the scale measured consistently with each other. Several validity investigations were also made in this study. It has been stated that the scales based on observers' ratings do not show strong correlations with the scales based on the patient's self-report, and these different types of scales evaluating different parameter may give conflicting results and sometimes compete with each other.[15] However, there are some reports stating high correlation between these two types of scales in the rating of depression.[16] HDRS scores showed a significant but moderate correlation (.48) with BDI scores in our study. This result may arise from the difference between the scales as mentioned above, or from the differences of items of BDI and HDRS. BDI items are mainly about cognitive properties and HDRS items are mainly physiological and general psychological symptoms. Another self-report scale, the Zung Depression Scale correlates moderately (.40 to .56) with the HDRS in two different studies.[17,18]

The HDRS also shows a moderate and significant correlation ($r = .56$; $P < .05$) with CGI scores. This correlation coefficient forms the basis for the testing of the validity of the similar scales. Some studies have reported similar correlations between HDRS, CGI, and BDI scores.[19-21] It is generally accepted that the self-report type of depression scales are easier to use, because they do not require an independent rater. However, according to some authors, observer-rated scales such as the HDRS are better than the self-rated scales to determine the severity of depression.[21] We also analyzed HDRS-BDI correlations in mild, moderate, and severe depressive groups in terms of CGI ratings. The correlation coefficient reached its highest value in the group with moderate depression (.66). In the two other groups, this value is as low as .28. However, it must be remembered that the number of subgroups is quite limited (11 to 14 subjects). HDRS-BDI correlations are high, except in severe disease states.[18]

The DSM-III-R has a subclassification coding based on severity of depression (1, mild; 2, moderate; 3, severe without psychotic features; 4, severe with mood congruent psychotic features; 5, severe with mood incongruent psychotic features). We found a significant but relatively low correlation of .37 between HDRS and DSM-III-R severity ratings. This low correlation may stem from the different classification criteria.

To investigate the structural validity, HDRS items were grouped in six factors on factor analysis. We named these factors according to their content as follows: "agitated," "anxious," "vegetative type," "somatized," and "retarded"; however, the last factor could not be named since it did not exert a significant composition. The first three factors were obtained from the original HDRS[1] (retarded, agitated and depression with anxiety), and we used the same names because these factors show the same weight as in the original study. In other words, these similarities indicate that the Turkish version and the original HDRS are concordant to a great extent, and that the Turkish version has structural validity. The correlation ($-.13$) between the controls and patient groups indicates that that the HDRS provides a good assessment of depression.

CONCLUSION

According to our study, the symptoms of depression are significantly more frequent in females compared with males, and educational level, marital status, and monthly income do not exert any effect on depression scores. The Turkish version of the HDRS has sufficient internal consistency; split-half, test-retest, and inter-rater reliability; structural and similar scales validity; and was shown to be valid and reliable in the assessment of clinical depression. We believe the HDRS is superior to the BDI in the evaluation of the severity of depression. To determine the features and differences of these two scales, new research should be done in larger samples having different characteristics.

REFERENCES

1. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56-62.
2. Anastasi A. Psychological Testing. Ed. 5. New York, NY: Macmillan, 1982.
3. Özgüven IE. Psychological Testing. Ankara, Turkey: Yeni Dogus, 1994.
4. Williams BW. A structured interview guide for Hamilton Depression Rating Scale. Arch Gen Psychiatry 1978;45:742-747.
5. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Ed. 3. Rev (DSM-III-R). Washington, DC: American Psychiatric Association, 1987.
6. Sorias S, Saygili R, Elbi H, Vahip S, Mete L, Nigirme Z, et al. DSM-III-R Structured Clinical Interview Turkish Version: SCID. Izmir, Turkey: Ege University Press, 1988.
7. Spitzer RL, Williams JBW, Gibbon M. Clinical Interview for DSM-III-R. New York, NY: New York State Psychiatric Institute, Biometrics Research Department, 1987.
8. Beneke M, Rasmus W. Clinical global impressions (ECDEU): some critical comments. Pharmacopsychiatrica 1992;25:171-176.
9. Beck AT. Depression: Clinical, Experimental, and Theoretical Aspects. Ed. 1. New York, NY: Hoeber Medical Division, Harper & Row, 1967.
10. Tegin B. Cognitive process in depression. Turk J Psychol 1987;6:116-121.
11. Snaith RP. Present use of the Hamilton Depression Rating Scale: observation on method of assessment in research of depressive disorder. Br J Psychiatry 1996;168:594-597.
12. Sercan M, Yüksel S. Somatic symptoms dominance in depressive disorders. Turk J Psychiatry 1990;1:2-7.
13. Cicchetti DV, Prusoff BA. Reliability of depression and associated clinical symptoms. Arch Gen Psychiatry 1983;40:987-990.
14. Senra C, Polaino A. Concordance between clinical and self-report depression scales during the acute phase and after treatment. J Affect Disord 1993;27:13-20.
15. Hamilton M. Mood disorders: clinical features. In: Kaplan HI, Freedman AM, Sadock BJ (eds): Comprehensive Textbook of Psychiatry. Vol. 5. Baltimore, MD: Williams & Wilkins, 1989:892-913.
16. Shain BN, Naylor M, Alessi N. Comparison of self-rated and clinician-rated measures of depression in adolescents. Am J Psychiatry 1990;147:793-795.
17. Grebb JA. Psychiatric rating scales. In: Kaplan HI, Freedman AM, Sadock BJ (eds): Comprehensive Textbook of Psychiatry. Vol. 1. Baltimore, MD: Williams & Wilkins, 1989:534-536.
18. Zung WKK. A cross-cultural survey of symptoms in depression. Am J Psychiatry 1969;126:116-121.
19. Bech P, Gram LF, Dein E, Jacobsen O, Vitger J, Bolwig TG. Quantitative rating of depressive states. Acta Psychiatr Scand 1975;51:161-170.
20. Bailey J, Coppen A. A comparison between the Hamilton Rating Scale and the Beck Inventory in the measurement of depression. Br J Psychiatry 1976;128:486-489.
21. Prusoff BA, Klerman GL, Paykel ES. Concordance between clinical assessments and patient's self report in depression. Arch Gen Psychiatry 1972;26:546-552.