Yale Depressive Symptom Inventory

Reliability and Validity of the Symptoms of Major Depressive Illness

Carolyn Mazure, PhD; J. Craig Nelson, MD; Lawrence H. Price, MD

• **In two consecutive studies, we examined the interrater reliability and then the concurrent validity of interview ratings for individual symptoms of major depressive illness. The concurrent validity of symptoms was determined by assessing the degree to which symptoms observed or reported during an interview were observed in daily behavior. Results indicated that most signs and symptoms of major depression and melancholia can be reliably rated by clinicians during a semi-structured interview. Ratings of observable symptoms (signs) assessed during the interview were valid indicators of dysfunction observed in daily behavior. Several but not all ratings based on patient report of symptoms were at variance with observation. These discordant patient-reported symptoms may have value as subjective reports but were not accurate descriptions of observed dysfunction.**

(*Arch Gen Psychiatry* 1986;43:451-456)

Approximately three dozen symptoms repeatedly have been associated with major depressive illness. These symptoms have been used in varying combinations to diagnose depression, to identify diagnostic subtypes in epidemiologic, genetic, and biologic studies of depression, and to assess severity and change during the treatment of depression. Although the use of biologic markers has been suggested as an alternate strategy to symptom-based systems for diagnosing depression, none of the currently available indexes has proved to be both adequately sensitive and specific. Thus, symptoms remain the primary basis for diagnosis and the primary measure of treatment response in outcome studies and in drug trials. As a consequence of the importance of symptom-based diagnosis and assessment, it is essential to establish that the rating of a symptom is reliable and valid.

Reliability of clinical measures in depression generally has been studied as interrater reliability. The focus of previous studies has usually been the overall reliability of an assessment instrument.[1-4] Few studies have examined the reliability of individual symptoms used for establishing diagnosis or rating severity and change.[5,6] Yet the interrater reliability of individual symptoms is particularly important because the usefulness of any symptom-based system depends on whether the meaning of each symptom can be agreed on and because the subtyping of major depression frequently depends on the reliable assessment of a small number of key symptoms.

While validity of clinical measures of depression has been less frequently studied, demonstration of validity is fundamental to establishing that a symptom rating or scale is meaningful in the context in which it is used. The issue is complicated by the fact that different forms of validity address different functions of a clinical measure and that establishing one type of validity does not establish other types. The validity of clinical measures of depression usually has been examined by studying agreement between two different depression rating instruments, such as the Hamilton Depression Rating Scale and the Beck Depression Inventory,[7] or the Hamilton scale and the Zung Self-rating Scale.[8] Although global agreement between two different rating scales establishes that each scale measures the same psychiatric construct, this does not address the validity of the individual items that make up the scale, nor does this establish how the scale relates to an independent or "criterion" measure of what is being assessed. Criterion-related validities provide data on what may be inferred from a clinical measure by relating it to another variable. A predictive relationship could be established between a clinical measure and a criterion if, for example, symptom ratings are highly correlated with some future outcome. A relationship between a clinical measure and a concurrent biologic variable suggests the biologic state that can be inferred from the clinical measure.

Although it is important to establish how a clinical measure relates to other variables, this approach to validity does not address the basic issue of whether a clinical rating measures what it purports to measure. When a patient reports a symptom, can we assume the symptom is present and that it might be observed in daily behavior outside the interview? Attempts have been made to relate self-reported

Accepted for publication Aug 19, 1985.

From the Department of Psychiatry, Yale University School of Medicine, New Haven, Conn.

Reprint requests to Yale-New Haven Hospital, MU 10-7, 20 York St, New Haven, CT 06504 (Dr Mazure).

depressive symptoms to interview ratings[9] and to relate interview ratings to clinical diagnosis[10] and to global clinical ratings.[3,11] Yet, to our knowledge, there has been no systematic study of the common individual symptoms of major depressive illness to determine the concurrent relationship between interview ratings or self-report ratings and observed daily behavior except for diurnal variation[12] and sleep.[13] Demonstration of this type of validity is crucial (1) if clinicians and investigators assume that symptoms rated during an interview are indeed present and (2) in basic research on the pathophysiology of depression for which an accurate definition of disordered behavior is essential. This last issue may become increasingly important as attempts are made to relate biologic dysfunction to disordered behavior.

The assessment of depressed behavior is complicated by the lack of any agreed-on "true measure" of a symptom. In addition, the nature of the symptoms to be measured varies from items that are purely observed signs to those that are patient reported. Patient-reported symptoms include reports of subjective experience as well as reports of observable behavior occurring outside the interview, eg, eating and sleeping. In the current study, we used direct observation of behavior as the validating criterion for symptoms because direct clinical observation has inherent validity, at least for those symptoms that have observable manifestations. We chose an observation period of three days to determine if symptoms were persistent and frequent.

Although it might be expected that signs observed in the interview would be more observable on the ward, one cannot assume that observation of behavior during a brief interview will necessarily be representative of patients' behavior outside the interview. We used nursing staff for the direct behavioral assessment because nurses observe patients over a 24-hour period in both formal and informal settings. This minimizes the possibility of the assessment process influencing the behavior assessed.[14] While we recognized that subjective symptoms might be more difficult to observe, many patients complain of their subjective distress, and those complaints can be observed and rated. Observations of other patient-reported symptoms, eg, eating and sleeping, could help to determine if patients' reports are an accurate description of what others observe.

We undertook this study to determine the reliability and validity of commonly assessed symptoms of depression such as the *DSM-III* symptom criteria for major depression and melancholia[15] and the items on the Hamilton Depression Rating Scale.[1,16] First, interrater reliability of individual items was determined using trained clinician-raters. The concurrent validity of individual symptoms then was assessed. Specifically, we examined the concordance between symptoms rated during a semistructured interview and observed behavior during a concurrent three-day period. For most items, such as appetite, we considered direct observation of the related behavior (eating) over a period of time as the best measure, or criterion, for the behavior. For other items, such as guilt, we utilized observation of the item as it was reflected in patients' everyday conversations and interactions. Finally, we examined the extent to which patients overreported or underreported the presence of symptoms in comparison with observation-based ratings of those same symptoms.

METHODS

Yale–New Haven Hospital Depression Symptom Inventory

A symptom inventory was compiled at Yale–New Haven (Conn) Hospital that (1) included *DSM-III* criteria for the diagnoses of major depression and melancholia, (2) focused on current symptoms, (3) incorporated items from the modified 24-item Hamilton Depression Rating Scale useful for sequential assessment of severity, (4) provided symptom ratings on the basis of clinician observation to the extent possible, and (5) used anchoring points appropriate for depressed inpatients. At the start of this project, no existing scale performed all these functions. Table 1 shows the symptoms constituting the Yale–New Haven Hospital Depressive Symptom Inventory. We included tension, as defined in the Brief Psychiatric Rating Scale (BPRS),[17] and ruminative thinking, which we have found useful as a diagnostic criterion for melancholia.[18,19] Several standard items were divided to allow greater specificity of symptom assessment: work and interest, suicidal ideation and suicidal behavior, and speech and motor retardation. The total number of symptoms was 34. All symptoms were rated on either a three-point (0 to 2) or five-point (0 to 4) scale. Global severity was assessed on a seven-point scale (0 to 6).

The symptom inventory was organized as a semistructured interview that emphasized observation of symptoms rather than patient self-report. Each item on the symptom inventory was based on (*a*) observation of symptoms during the interview, (*b*) both patient report and observation, or (*c*) patient report only. Items were considered both reported and observed if the patient could be directly questioned about the item and/or the rater could observe whether the patient spontaneously referred to the particular concern (eg, somatic concern, hopelessness) throughout the interview. These items differed from symptoms rated solely on the basis of patients' reports because they occurred outside the interview (eg, diurnal variation or decreased sleep) or because they could not be observed (eg, decreased sexual interest).

Interrater Reliability

To determine interrater reliability, three clinicians simultaneously rated 31 depressed inpatients (26 women and five men), with each clinician alternating as interviewer. Consecutively admitted patients who had *DSM-III* diagnoses of major depression and who were nonpsychotic were interviewed, generating a sample of patients having depression of variable severity. We elected to exclude psychotic depressive patients since we were concerned that the psychosis itself would increase the distortion of patient-reported information. Patients ranged in age from 18 to 83 years (mean ± SD, 45.3 ± 17 years). Two of the female patients had a diagnosis of bipolar disorder but were depressed at the time of the interview. The average time in the hospital from admission to time of interview was 12 ± 11 days.

Concurrent Validity

To determine concordance, a second independent sample of 31 consecutively admitted patients who met *DSM-III* criteria for major depressive episode and were not psychotic was obtained. These included 24 female and seven male patients ranging in age from 16 to 82 years (mean, 53 ± 17 years). Two of the female patients in this sample had bipolar disorder and two others had probable diagnoses of bipolar disorder, but all were depressed at the time of the study.

Validity

To establish behavioral measures with which the validity of the interview items could be tested, a 26-item nursing observation checklist analogous to the clinician-rated symptom inventory was constructed. Each nursing checklist item had the same rating scale and anchor points as the symptom inventory. This checklist was completed by one nurse with information obtained from nursing shifts over a three-day period. Nursing staff did not perform a structured interview, nor did they ask a series of questions about symptoms. They were asked to observe patients' behavior with knowledge of the symptoms being assessed. In addition to observing physical and interpersonal behavior, they were instructed to observe the patients' verbal behavior, particularly noting spontaneous reports, complaints, and themes of conversations; for example, if a patient continually spoke of feeling worthless and a failure, then we rated worthlessness as present. These items were rated on manifest content, not what the themes might symbolically represent. Sleep was rated using sleep checks at 30-minute inter-

Table 1.—Yale–New Haven (Conn) Hospital Depressive Symptom Inventory

Item	Source of Item: Hamilton Depression Rating Scale	*DSM-III* Major Depression Criteria	*DSM-III* Melancholia Criteria
Depressed mood	X	X	...
Distinct quality of mood	...	...	X
Psychic anxiety	X	...	...
Somatic anxiety	X	...	...
Difficulty falling asleep	X	X	...
Middle of night awakening	X	X	...
Early-morning awakening	X	X	X
Loss of appetite	X	X	X
Loss of weight	X	X	X
Loss of energy	X	X	...
Decreased interest	X	X	...
Loss of pleasure	...	...	X
Decreased work	X	...	...
Decreased sexual interest	X	X	...
Decreased concentration	...	X	...
Diurnal variation, AM	X	...	X
Diurnal variation, PM	X	...	...
Depersonalization	X	...	...
Obsessional symptoms	X	...	...
Ruminative thinking	...	...	...
Paranoid symptoms	X	...	...
Guilt	X	X	X
Worthlessness	X	X	...
Helplessness	X	...	...
Hopelessness	X	...	...
Suicidal ideation	X	X	...
Suicidal behavior	X	X	...
Somatic concern	X	...	...
Loss of insight	X	...	...
Speech retardation	X	X	X
Motor retardation	X	X	X
Agitation	X	X	X
Tension	...	...	...
Lack of responsiveness	...	...	X

vals throughout the night. Table 2 shows the partition of symptoms according to the source and type of rating.

Of the 34 symptoms on the symptom inventory, eight were not included in the validity study. A symptom was excluded if (1) it could not be reliably rated (lack of insight, tension), (2) it was not possible to define an observable manifestation of the symptom (distinct quality of mood), (3) the symptom was actively prevented in the hospital (suicidal behavior, weight loss), or (4) it was infrequent in nonpsychotic patients with major depression, ie, present in less than 20% of the patients (depersonalization, obsessional symptoms, and paranoia). Thus, 26 items plus global severity constituted the list of 27 behavioral assessments that were compared with ratings made during the clinical interview using the symptom inventory.

The nursing observation checklist was performed the same day as the Yale–New Haven Hospital Depressive Symptom Inventory. The three-day period of observation for the nurses' checklist was concurrent with the period described by patient report in the interview since patients were asked to describe symptoms present during the days prior to the interview. The interview and observation checklist were performed an average of 16 days following admission. Throughout the study, one nurse compiled all the nursing observations, and the clinician completing the symptom inventory ratings also remained the same. These two raters were blind to each others' ratings.

RESULTS

Interrater Reliability

Interrater reliability was examined for 30 depressive symptoms. Table 3 presents the intraclass correlations[20,21] for the three clinician-raters on 28 symptoms plus the measure of global severity. Twenty-four symptoms and assessment of global severity showed high intraclass correlations, with $\alpha = .001$. Two items, speech retardation and motor retardation, showed moderate reliability across three trained raters. Correlations were low for agitation and for tension. On closer examination, it appeared that agitation was rarely rated above a scale score of 1 in this patient group, resulting in a restricted range for this rating and compromising our determination of the reliability of the item. For tension, the low correlation reflected disagreement among raters or lack of reliability for the rating.

Two symptoms, distinct quality of mood and insight, were analyzed using a κ coefficient[22] rather than by intraclass correlation because each of these items was a dichotomous variable. Distinct quality of mood was reliable across all three raters, but lack of insight was not.

In sum, results indicated that two symptoms, tension and lack of insight, could not be reliably rated. One symptom, agitation, was present only at a low level of severity, which limited its analysis. Interrater reliability could not be adequately assessed in four infrequent symptoms (depersonalization, obsessional symptoms, paranoid symptoms, and suicidal behavior).

Concordance

Of 26 symptoms studied, three (diurnal AM worsening, PM worsening, and decreased sexual interest) were rarely observed by nursing staff. These three symptoms were occasionally reported during the clinical interview; however, diurnal variation was not reflected in observed behavior and sexual interest was not reflected in conversation.

Table 4 shows the concordance using Pearson's product-moment correlations (r) between clinician ratings and nursing observations, with clinician-rated items categorized according to the type of rating (clinical observation during interview only, both report and observation during interview, or patient report only). We considered correlations with probability levels above .01 as indicating low concordance. Discordant items had nonsignificant correlations ($P > .05$). The data indicated that all depressive symptoms based entirely on observation during the interview (signs) were accurate reflections of daily behavior. Of the eight symptoms obtained by combined patient report and observation, six were concordant and two (worthlessness and helplessness) showed low concordance or discordance, respectively. Although some patient-reported symptoms were reflected in behavior outside the interview, most did not adequately correlate with observations.

Nine symptoms were discordant or had low concordance. A Wilcoxon matched-pairs signed-ranks test[23] was used to examine the direction and magnitude of disagreement between interview and ward observation for these items. For four of these nine symptoms, the ratings were significantly different (two-tailed test). Patients underrated helplessness ($P < .05$) and psychic anxiety ($P < .05$) and overrated difficulty falling asleep ($P < .05$) and

Table 2.—Source and Type of Ratings

Interview Ratings	Nurses' Ratings of Ward Behavior: Observed Only	Observed and Reported	Patient Reported or Included in Conversation
Observed only	Lack of responsiveness; motor retardation; speech retardation; agitation	...	...
Observed and reported	...	Depressed mood; ruminative thinking	...
Patient reported or included in conversation	...	...	Hopelessness; helplessness; worthlessness; guilt; somatic concern
Patient reported	Appetite/eating; difficulty falling asleep; middle of night awakening; early-morning awakening; AM worsening; PM worsening	Decreased concentration; loss of pleasure; loss of interest; decreased work; psychic anxiety; somatic anxiety	Suicidal ideation; loss of energy; decreased sexual interest

middle of the night awakening ($P<.01$). The remaining five items demonstrated no consistent pattern for the lack of agreement between patient report and observation.

COMMENT

Our findings indicate that most symptoms of major depressive illness can be reliably rated by clinicians during an interview whether those assessments are based on observation of signs or on patients' reports of symptoms.

Establishing validity of depressive symptoms is more complicated and depends in part on whether a symptom can be directly observed. The four symptoms rated on the basis of observation only, and six of the eight symptoms based on observation and patient report, were significantly correlated with behavior seen outside the interview. However, of the 12 interview assessments based solely on patients' reports of symptoms, only five were significantly correlated with observation of daily behavior.

It might be expected that observed signs would be more likely to correlate with behavior seen outside the interview. However, one cannot assume that behavior observed in the interview would necessarily be observed on the ward. One symptom, lack of responsiveness, was designed to assess lack of reactivity through observation. This differed from loss of pleasure, which was rated on the basis of the patients' reports but was related to the same underlying phenomenon of anhedonia. At the time the study was initiated, we were not certain if observation during a brief interview would prove to be a valid indicator of responsiveness as observed on the ward. In fact, this symptom had the highest concordance of any rating including loss of pleasure. Ruminative thinking, which we described in greater detail elsewhere,[19] was another new item that we tested in this study and found to be a valid indicator of observed behavior.

The possibility that we are merely finding that observed behavior can be observed and subjective symptoms cannot does not explain the specific findings of this study. While symptoms rated on the basis of patients' reports during the interview were in general less likely to be validated, five of the 12 such symptoms were concurrent with observable manifestations of the symptoms. Among those symptoms that were validated, some were more easily observed, eg, appetite/eating and decreased work, but others were more subjective, eg, loss of pleasure, loss of interest, and suicidal thinking.

Other patient-reported symptoms were at variance with observational reports. These discordant items included psychic anxiety, somatic anxiety, and three *DSM-III* symptom criteria—lack of energy, impaired concentration, and decreased sleep. Three of these items (psychic anxiety, somatic anxiety, and lack of energy) were primarily subjective symptoms and the question arose whether ward observation was the best measure for these items. While agreement between the interview ratings and the ward observations would have supported the validity of these symptoms, our failure to find agreement does not necessarily mean that interview ratings of these symptoms are invalid. It is possible that patients' reports for these discordant items may have value as a subjective account even though the data indicate that these patients' reports were not accurate indicators of observable manifestations of those symptoms.

Discordance of items was not explained simply on the basis that the patients overrated the symptoms. For example, psychic anxiety and helplessness were rated as more severe by nurses than by patients. Furthermore, neither the concordant nor the discordant items appeared to be limited to one sphere of functioning. For example, patients were at variance with observation of neurovegetative changes in terms of sleep but not in terms of eating.

Discordant, Untested, and Divided Symptoms

Impaired concentration conceivably was observable and thus measurable. However, opportunities to observe decreased concentration in our depressed inpatients might have been limited because of diminished involvement in tasks requiring sustained and thus observable concentration.

Sleep or lack of sleep was directly observed at 30-minute intervals throughout the night. These observed ratings seemed an accurate criterion measure of the sleep disturbance symptoms as defined. Although electroencephalographic sleep recordings give a precise measure of sleep and sleep stage, we were interested in whether patients were asleep or not. Furthermore, this item did not address the more subjective question of whether patients felt "rested." The lack of agreement between patients' reports and observations of sleep suggests that patients' reports are not valid descriptions of sleep disturbance. This differs from a previ-

Table 3.—Interrater Reliability for 30 Symptoms

Symptom	Intraclass Correlation Coefficient*
Weight loss since onset of episode	.98
Decreased appetite	.94
Psychic anxiety	.90
Somatic anxiety	.90
Suicidal ideation	.90
Loss of interest	.88
Worthlessness	.86
Helplessness	.86
Loss of energy	.85
Guilt	.85
Diurnal worsening: PM	.84
Somatic concern	.83
Decreased work	.81
Difficulty falling asleep	.80
Early-morning awakening	.80
Distinct quality†	. . .
Ruminative thinking	.78
Diurnal worsening: AM	.79
Hopelessness	.79
Middle of night awakening	.77
Loss of pleasure	.74
Decreased sexual interest	.73
Decreased concentration	.71
Depressed mood	.70
Lack of responsiveness	.67
Speech retardation	.59
Motor retardation	.57
Lack of insight‡	. . .
Agitation	.40
Tension	.30
Global severity	.80

*$P<.001$.

†κ values were as follows: rater 1 vs rater 2, 0.80; rater 1 vs rater 3, 0.79; and rater 2 vs rater 3, 0.75.

‡κ values were as follows: rater 1 vs rater 2, 0.73; rater 1 vs rater 3, 0.31; and rater 2 vs rater 3, 0.31 (restricted range for rater 3).

Table 4.—Concordance (r) of Three Types of Interview Ratings With Nursing Observations

Basis of Interview Rating	Symptom	Correlation With Nursing Observation
Clinical observation only	Lack of responsiveness	.696*
	Motor retardation	.528†
	Agitation	.476†
	Speech retardation	.471†
Both patient report and clinical observations	Global severity	.598*
	Hopelessness	.580*
	Depressed mood	.578*
	Ruminative thinking	.547†
	Somatic concern	.502†
	Guilt	.465†
	Worthlessness	.402‡
	Helplessness	.264
Patient report only	Decreased appetite	.659*
	Loss of pleasure	.512†
	Suicidal ideation	.499†
	Loss of interest	.495†
	Decreased work	.495†
	Middle of night awakening	.411‡
	Psychic anxiety	.385‡
	Early-morning awakening	.337
	Somatic anxiety	.278
	Loss of energy	.275
	Decreased concentration	.148
	Difficulty falling asleep	.100

*$P<.001$.
†$P<.01$.
‡$P<.05$.

ous observation,[18] but the difference in findings may relate to the range of disturbance observed. It may be possible to discriminate between sound sleep and very disrupted sleep, but we previously reported that most inpatients (67% to 83%) with major depression have some degree of sleep disturbance.[24,25] Our data suggest that these patients do not accurately report the severity of sleep disturbance. This finding has direct implications for use of sleep items as diagnostic criteria. The *DSM-III* lists early-morning awakening as a symptom criterion for melancholia and defines it as awakening two hours before the usual time. Our data suggest that patients cannot rate sleep with this degree of accuracy, and if patients' reports are used to fulfill this criterion, the item should be viewed as an approximation.

The reliability of one *DSM-III* symptom of major depression, agitation, could not be adequately assessed because the range of this item was restricted. Exclusion of psychotic depressive subjects may have contributed to this finding. It is also possible that the method of patient selection may have affected this finding since severely agitated patients could not have participated in the interview. Similarly, concurrent validity could not be examined for the *DSM-III* symptoms of decreased sexual interest, diurnal AM worsening, and diurnal PM worsening because these three symptoms were not observed in this sample. We noted previously[26] that patients frequently denied having sexual interest while in the hospital and away from their partners. Another study[12] of nine unipolar depressed patients reporting diurnal variation also found that self-report of fluctuation in mood was not verified in behavioral observation. While some individual patients on our unit have had observable diurnal variation, it is our impression that among nonpsychotic patients, the frequency is low. None was included in this sample. Distinct quality of mood was not assessed because a criterion measure could not be determined, and weight loss was not assessed because we intervened to prevent it.

Examination of the usefulness of dividing symptoms into two component parts showed that suicidal thinking and behavior had different frequencies of occurrence in our currently hospitalized patients. Suicidal behavior was prevented in the hospital, yet suicidal thinking persisted. Consequently, it seemed that two different symptoms were represented and it was useful to measure them separately. The correlation of interview ratings with observation ratings for speech retardation (.471) and motor retardation (.528) was slightly improved (.660) when the assessments for the two items were summed. This was also the case for work (.495) and interest (.495) when combined (.597). Separating each of these items into two symptoms may decrease somewhat the concordance of interview and observational ratings.

Validity: Importance and Type

The type of validity a symptom possesses determines how it can be used. We have examined the concurrent relationship between a test measure (eg, a symptom rating) and

a criterion (eg, behavior itself). This type of validity is important when the description of the actual individual symptom is in question. For example, to understand the basic pathophysiology of depression, it is important to know if early-morning awakening really exists. The current data suggest that patients' reports of early-morning awakening cannot be used as a valid description of this symptom. However, patients' reports of the same symptom may be valid in a different context or for different purposes.

In a previous study[26] of drug-responsive symptoms in melancholia, we determined that certain depressive symptoms varied in direct relation to therapeutic desipramine plasma concentrations. Four of the ten symptoms that changed in direct response to medication were among the discordant symptoms in the present study. These four symptoms (lack of energy, worthlessness, somatic anxiety, and early-morning awakening) appeared to be valid indicators of drug response, although in this study the correlation of report with observed behavior was not established. These symptoms may be useful perceptions of subjective distress that is drug responsive but not accurate indicators of observable behavior.

Examination of individual symptoms is essential to establishing confidence in symptom-based diagnosis and assessment. Even if global severity and change are measured, such measurements are based on composite judgments of individual symptoms. The present data indicate that most symptoms of depression can be reliably rated. Those that can be observed during a clinical interview are likely to be validated by observation of daily behavior, a finding that supports the use of clinical observation in symptom assessment. The validity of patient-reported symptoms is more variable. Some patient-reported symptoms can be validated by observation, are reliable, and appear to be useful in diagnosis and treatment of depression. Others were not validated, but the reason may vary depending on the nature of the symptom. Some patient-reported symptoms primarily reflect subjective experience and may be difficult to validate using observation. Other symptoms, however, may be reported by patients because they occur outside the interview but are observable when they occur. For these symptoms, lack of validation suggests that the patients' reports are not accurate. Ultimately, accurate use of patients' reports requires validation of those reports. If observation does not provide that validation, accuracy of symptom reports and assessment of actual dysfunction remain unclear.

This study was supported in part by grant 1 P50 MH30929 from the National Institute of Mental Health, Bethesda, Md (to Yale Mental Health Clinical Research Center).

We would like to express our appreciation to Sarah Roumanis, RN, and the nursing staff of the Adult Treatment Unit for the help provided in conducting this study.

References

1. Hamilton M: A rating scale for depression. *J Neurol Neurosurg Psychiatry* 1960;23:56-62.
2. Montgomery SA, Asberg M: A new depression scale designed to be sensitive to change. *Br J Psychiatry* 1979;134:382-389.
3. Knesevich JW, Biggs JT, Clayton PJ, Ziegler VE: Validity of the Hamilton Depression Rating Scale. *Br J Psychiatry* 1977;131:49-52.
4. Mazure C, Gershon ES: Blindness and reliability in lifetime psychiatric diagnosis. *Arch Gen Psychiatry* 1979;36:521-525.
5. Cicchetti DV, Prusoff BA: Reliability of depression and associated clinical symptoms. *Arch Gen Psychiatry* 1983;40:987-990.
6. Paykel ES: The clinical interview for depression: Development, reliability, and validity. *J Affective Disord* 1985;9:85-96.
7. 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.
8. Carroll BJ, Fielding JM, Blashki TG: Depression rating scales: A critical review. *Arch Gen Psychiatry* 1973;28:361-366.
9. Prusoff BA, Klerman GL, Paykel ES: Concordance between clinical assessment and patients' self-report in depression. *Arch Gen Psychiatry* 1972;26:546-552.
10. Thase ME, Hersen M, Bellack AS, Himmelhoch JM, Kupfer DJ: Validation of a Hamilton subscale for endogenomorphic depression. *J Affective Disord* 1983;5:267-278.
11. Bech P, Gram LF, Dein E, Jacobsen O, Vitger J, Bolwig TG: Quantitative rating of depressive states. *Acta Psychiatr Scand* 1975;51:161-170.
12. Williams JG, Barlow DH, Agras WS: Diurnal variation in depression: Is it there? *J Nerv Ment Dis* 1975;161:59-62.
13. Costello CG, Selby MM: The relationship between sleep patterns and reactive and endogenous depressions. *Br J Psychiatry* 1965;111:497-501.
14. Bunney WE, Hamburg DA: Methods for reliable longitudinal observation of behavior. *Arch Gen Psychiatry* 1963;9:280-294.
15. American Psychiatric Association Committee on Nomenclature and Statistics: *Diagnostic and Statistical Manual of Mental Disorders*, ed 3. Washington, DC, American Psychiatric Association, 1980.
16. Hamilton M: Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967;6:278-296.
17. Overall JE, Gorham DR: The Brief Psychiatric Rating Scale. *Psychol Rep* 1962;10:799-812.
18. Nelson JC, Charney MD, Quinlan DM: Characteristics of autonomous depression. *J Nerv Ment Dis* 1980;168:637-643.
19. Nelson JC, Mazure C: Ruminative thinking: A distinctive sign of melancholia. *J Affective Disord* 1985;9:41-46.
20. Bartko JJ: The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966;19:3-11.
21. Bartko JJ: On various intraclass correlation reliability coefficients. *Psychol Bull* 1976;83:762-765.
22. Cohen J: A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37-46.
23. Siegel S: *Nonparametric Statistics*. New York, McGraw-Hill Book Co, 1956.
24. Nelson JC, Charney DS: Primary affective disorder criteria and the endogenous-reactive distinction. *Arch Gen Psychiatry* 1980;37:787-793.
25. Nelson JC, Charney DS, Quinlan DM: Evaluation of the *DSM-III* criteria for melancholia. *Arch Gen Psychiatry* 1981;38:555-559.
26. Nelson JC, Mazure C, Quinlan DM, Jatlow PI: Drug-responsive symptoms in melancholia. *Arch Gen Psychiatry* 1984;41:663-668.