Psychological Assessment:
A Journal of Consulting and Clinical Psychology
1990, Vol. 2, No. 2, 134–138

1040-3590/90/$00.75

Depression Screening Scale Validation in an Elderly, Community-Dwelling Population

Karl W. Stukenberg
Ohio State University

Jason R. Dura and Janice K. Kiecolt-Glaser
Department of Psychiatry
Ohio State University College of Medicine

This study contrasted the relative effectiveness of an interviewer-rated instrument, the Hamilton Depression Rating Scale, and 2 self-report scales, the short form of the Beck Depression Inventory and the depression scale from the Brief Symptom Inventory, in identifying cases of depression. Cases of major depression, dysthymia, and depressive disorder not otherwise specified (NOS) were identified by means of the Structured Clinical Interview for *DSM–III–R* (SCID) in a sample of 177 elderly community-dwellers. Receiver operating curves were used to evaluate the relative abilities of the 3 screening instruments to identify cases of depression. All 3 instruments identified major depression and depressive disorder NOS. None was consistently sensitive to cases of dysthymia. The incremental utility of the interview-based instrument for screening was nonsignificant, suggesting that the increased expense in a community setting may not be justified.

Depression screening instruments have been designed to identify potential cases of depression quickly, inexpensively, and with a minimum of psychometric time or training. Paper-and-pencil inventories meet these criteria extremely well (Yesavage et al., 1983). Interviewer-rated depression rating scales, although more expensive and time-consuming, allow for normatively based symptom-severity comparisons on the part of a trained clinician (Hamilton, 1967). In the present study, we evaluated the incremental utility of an interview-based screening scale in addition to pencil-and-paper scales for an elderly, community-dwelling population.

Depression scales, whether self-report or interviewer-rated, provide continuous measures of symptomatology. In practice, however, whether in the clinic or in the research lab, they are used to identify potential cases of depression. The transformation of a continuous measure to a discrete measure is generally quite simple: a cutoff score is created. The ability of the cutoff score to predict the presence or absence of a case must then be determined.

A number of techniques have been used in the evaluation of psychological screening instruments' validity, with clinical diagnosis as a criterion. Measures of sensitivity (or ability to detect cases of depression) and specificity (or ability to correctly identify people *not* experiencing depression) are generally reported for one, two, or three cutoff scores. Receiver operating characteristics (ROC) curves, on the other hand, enable one to evaluate the relative performance of various screening measures at all possible cutoff points (Murphy et al., 1987). They have been used by medical researchers to assess the predictive ability of various tests: for example, to assess the ability of pathologists to predict disease from various types of images or to predict benefits of treatment (Hanley & McNeil, 1982; Metz, 1978).

Recently, researchers have begun to use ROC curves in psychological studies (Koenig, Meador, Cohen, & Blazer, 1988). Hanley and McNeil (1982) have suggested that the area under the ROC curve is equivalent to the probability that in a random pair of case/noncase subjects, the scores on an inventory will allow each of them to be correctly identified. In the current study, we used ROC curves to assess the relative abilities of the three depression rating scales to predict depression in an elderly, community-dwelling population.

Depression is a significant problem among the community-dwelling elderly (Blazer & Williams, 1982) and perhaps particularly so among the chronically stressed (George & Gwyther, 1986; Kiecolt-Glaser, Dyer, & Shuttleworth, 1988). The depressed elderly pose a problem for the health community because they often present themselves to the primary medical care system for treatment and assessment of mental health disorders (Redick & Taube, 1980). However, studies with middle-aged and older adults suggest that nonpsychiatric physicians fail to detect most cases of depression (Koenig et al., 1988; Nielsen & Williams, 1980; Rapp, Parisi, Walsh, & Wallace, 1988).

Structured diagnostic interviews provide a reliable means of identifying cases defined by the *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed., rev.; *DSM–III–R;* American

This research was supported by Grant MH42096 from the National Institute of Mental Health and by funds from the Office of Geriatrics and Gerontology, Ohio State University College of Medicine.

The authors thank James N. Butcher and an anonymous reviewer for helpful comments on a draft of this article.

Correspondence concerning this article should be addressed to Karl W. Stukenberg, who is now in the Division of Psychology, Department of Psychiatry, Baylor College of Medicine, 1200 Moursund Avenue, Houston, Texas 77030.

Psychiatric Association, 1987) as depressive disorders, including major depression, dysthymia, and depression not otherwise specified (NOS; Riskind, Beck, Berchick, Brown, & Steer, 1987). Although structured interviews can provide more reliable diagnoses, the length of the interviews, as well as their complexity, precludes their use in a general medical practice on a routine basis (Norris, Gallagher, Wilson, & Winograd, 1987). However, the use of structured interviews to provide target criteria is warranted in the evaluation of screening instruments.

A number of pencil-and-paper instruments have been developed to screen for depression. Most were developed for use with middle-aged populations, and their use with the elderly requires empirical validation. The elderly have higher base rates for somatic symptoms, and reporting these symptoms can unrealistically inflate scores on somatically laden depression scales (Zemore & Eames, 1979). Thus, even though somatic symptomatology is an important aspect of the depressive syndrome among the elderly (Norris et al., 1987), somatically laden items are generally not included in gerontological depression scales to decrease the number of false positives.

Interviewer-based depression rating scales have characteristics of both the pencil-and-paper inventories and structured interview diagnoses. Like the pencil-and-paper inventories, interviewer-based depression rating scales provide a continuous measure based on the aggregate number and severity of symptoms. But, because the scale is administered by trained clinicians, the normatively based assessment of somatic symptoms is included in the aggregate score total, which is derived from clinical decisions concerning the temporal qualities of the somatic symptoms.

In a recent review of the literature, Gallagher (1986) called for more studies establishing the reliability and validity of the Beck Depression Inventory (BDI) among older adults. Pencil-and-paper inventories generally have been shown to have adequate screening capabilities when used with elderly medical patient groups (Magni, Schifano, & de Leo, 1986; Noris et al., 1987; Okimoto et al., 1982; Rapp et al., 1988). However, in a recent study, an interviewer-based depression scale was substantially more satisfactory in screening for physician-rated depression than was a paper-and-pencil inventory among a group of elderly, predominantly Black and Hispanic medical outpatients (Toner, Gurland, & Teresi, 1988).

In the present study, we assessed the relative abilities of two pencil-and-paper depression rating scales: the Beck Depression Inventory Short Form (Beck & Beck, 1972) and the Brief Symptom Inventory (BSI) depression scale (Derogatis & Spencer, 1982). We contrasted these with an interviewer-based depression rating scale, the Hamilton Depression Rating Scale (HDRS; Hamilton, 1967). These scales' ability to identify cases of depression among an elderly, community-dwelling population was assessed by means of ROC curves. We hypothesized that the HDRS would be more sensitive to the presence of depression among the elderly than both the BSI depression scale and the BDI, without increasing the rate of false positives. We assumed that all three would distinguish cases of current *DSM-III-R* depressive disorders from noncases at better than chance levels. Using the BDI as a benchmark, we were also interested in establishing the relative predictive power of the BSI depression scale on an elderly, community-dwelling population.

Method

Subjects

Subjects were 177 community-dwelling adults, all over age 55 years. Mean age was 67.40 years (SD = 7.20), with a range from 56 to 88 years. There were 59 men and 118 women. Of these, 111 subjects were caring for relatives with Alzheimer's disease, and 66 were sociodemographically comparable subjects without similar caregiving responsibilities. Caregivers were recruited through multiple sources, including neurologists, hospital dementia evaluation units, nursing homes, dementia caregiver support groups, respite care programs, and governmental caregiver support programs. Control subjects were recruited through churches and advertisements in local newspapers and area newsletters.

Subjects were drawn from a larger study addressing health, immune function, and chronic stress. As a result, subjects who were not taking immunosuppressive medications and whose health problems did not have an immunological component (e.g., potential subjects were screened for cancer or recent surgeries) were selected. All subjects were paid $30 for participation in the study.

Diagnosis

The Structured Clinical Interview for *DSM-III-R*, Non-Patient Version (SCID-NP: Spitzer, Williams, Endicott, & Gibbon, 1987) was used for the determination of lifetime and present psychiatric disorders based on *DSM-III-R* criteria. In accordance with Norris et al.'s (1987) procedures, somatic symptoms were disregarded when it was clear that they were directly attributable to physical problems or when symptoms were longstanding and temporally unrelated to depressive affect.

Three advanced graduate students in clinical psychology and a postdoctoral fellow in clinical psychology made current diagnoses of the mood disorders of major depression, dysthymia, and depressive disorder not otherwise specified (NOS). A 75-subject subsample interrater reliability study was conducted. When three primary interviewers were compared with a secondary interviewer, an overall kappa on presence of psychiatric disorders was .92, well above the satisfactory level of .70 suggested by Riskind et al. (1987).

Interview-Based Rating Scale

The Hamilton Depression Rating Scale (HDRS; Hamilton, 1967) was administered after the SCID-NP. This commonly used, 24-item scale relies on an interviewer to rate each of 24 depressive symptoms. Each symptom has a unique scale of severity, and an overall score is calculated by summing the ratings for each symptom. As was the case with the SCID-NP, symptoms were disregarded if they were clearly attributable to physical disorders or if they were longstanding and their onset was temporally unrelated to other depressive symptoms. Interrater reliability, computed on data from 30 subjects (17% of the sample), was .85.

Pencil-and-Paper Scales

The short form of the Beck Depression Inventory, a 13-item subset of the original instrument (Beck & Beck, 1972), was used. It is commonly used to screen for depression and has been used with older populations, in part because it has fewer somatically laden items than some more biologically based scales (Hammen, 1980). Each cluster of items presents four sentences, and the subjects are to endorse the sentence or sentences that best describe the way they have been feeling in the past week.

The BSI (Derogatis & Spencer, 1982), a 53-item general psychopathology instrument designed for use in screening medical populations, includes a 6-item Depression scale without somatically laden items.

Subjects are asked to indicate to what extent they were distressed by each of the symptoms during the past 7 days, including the day of the test. Responses are made on a 5-point Likert scale with anchors at *not at all* and *extremely*. The scale has been normed with an elderly population (Hale, Cochran, & Hedgepeth, 1984), but validation studies have not, to our knowledge, been performed with an elderly depressed population.

A 10-item version of the Marlowe-Crowne Social Desirability Scale (Strahan & Gerbasi, 1972) was also included to assess the relation between depressed symptoms, both reported and observed, and social desirability (Crowne & Marlowe, 1960). Unlike earlier social desirability scales, the content has minimal overlap with well-being scales (Kozma & Stones, 1987).

Results

By *DSM-III-R* diagnosis, 24 of the 178 subjects who completed the SCID were currently depressed. Eight of the subjects received a diagnosis of major depression; 7, a diagnosis of depressive disorder NOS; and 9, a diagnosis of dysthymia.

ROC curves for the BDI, the BSI Depression scale, and the HDRS are plotted in Figure 1. The sensitivity (or true-positive rate) and the specificity (or false-positive rate) are plotted for every score on each instrument. The estimated area under the HDRS ROC curve was .85 (SE = .05), the BDI ROC curve area estimate was .82 (SE = .06), and the estimated area under the BSI Depression scale ROC curve was .83 (SE = .04), as estimated by means of Hanley and McNeil's (1982) conservative calculations. All areas were significantly greater than the area of no information (an area of .50), $p < .001$, but there were no significant differences between estimated areas of the three curves. The area of the HDRS was not significantly greater than either the BDI (z = 0.67) or the BSI Depression scale curve (z = 0.35). The BSI Depression scale curve was not significantly greater than that of the BDI (z = 0.43). All comparisons were calculated with the estimates of average correlation provided by Hanley and McNeil (1983).

The BDI short form recommended cutoff scores for determining mild, moderate, and severe depression are 5, 8, and 16, respectively (Beck & Beck, 1972). When a cutoff score of 5 was used, the BDI correctly identified 74% of the cases (120 of 163). The sensitivity of the measure was 0.71 and the specificity was 0.83. When a cutoff score of 8 was used, the scale correctly identified 88% of the sample (144 of 163), with a sensitivity of 0.59 and a specificity of 0.93. With a cutoff score of 16, the BDI correctly identified 90% of the sample (147 of 163), with a sensitivity of 0.29 and a specificity of 0.99.

The BSI Depression scale produces normatively based T scores. At a cutoff of 60, when compared with the original normal sample (Derogatis & Spencer, 1982), the BSI Depression scale correctly identified 79% of the cases (157 of 198). The sensitivity of the measure was 0.76 and the specificity was 0.77. At a cutoff of 65, the BSI Depression scale correctly identified 87% of the sample (173 of 198), with a sensitivity of 0.43 and a specificity of 0.92. At a cutoff of 70, the BSI Depression scale correctly identified 91% of the sample (180 of 198), with a sensitivity of 0.29 and a specificity of 0.98.

The BDI scores for dysthymic cases ranged from 1 to 7, so that dysthymic cases were identified only at the lowest cutoff. Fewer than half of the BSI Depression scale scores for cases of dysthymia were above a T score of 60. The HDRS score for two cases of dysthymia was zero.

The correlation between the BDI short form and the BSI Depression scale was .71 ($p < .0001$; n = 145). The BDI correlated .68 ($p < .0001$; n = 146) with the HDRS. The correlation between the HDRS and the BSI Depression scale was .60 ($p < .0001$, n = 177).

The three rating scales were also correlated with a 10-item version of the Marlowe-Crowne Social Desirability Scale (Strahan & Gerbasi, 1972). The BDI correlated −.11 ($p > .10$; n = 143), the BSI Depression scale correlated −.16 ($p < .05$; n = 173), and the HDRS correlated .04 with the Marlowe-Crowne scale.

Discussion

The use of ROC curves allowed for the statistical comparison of the three scales. The area under the HDRS ROC curve was larger, though not significantly larger, than the areas under the BDI and the BSI Depression scale curves. Thus, the BDI and the BSI Depression scale were comparable to the HDRS in the ability to screen for cases of depression in an elderly, community-dwelling sample. The anticipated increment in sensitivity resulting from the use of an interviewer-rated scale was not significant. This suggests that the increased time and expense involved in administering interview screening scales does not seem to be warranted in a nonclinical setting.

Neither the self-report measures nor the Hamilton scale discriminates between syndromal depression due to affective disorders versus other disorders. Thus, the depressive syndrome occurring in the context of other nonaffective disorders (e.g., schizophrenia and organicity) would render a score similar to the affective disorder score. Although it is probable that a trained interviewer administering the Hamilton would be able to determine other causes of depression, it should be emphasized that scores on all of the screening scales suggest the possibility of an affective disorder, but only a diagnostic interview like the SCID can provide a definitive diagnosis.

The BDI and BSI depression scale had reasonable agreement with *DSM-III-R* major depression and depression NOS diagnoses, which is consistent with prior studies involving elderly medical patients. In elderly medical outpatient screening studies, researchers using the Self-Rating Depression Scale (SDS: Zung, 1965) and the BDI and the Geriatric Depression Scale (GDS: Yeasavage et al., 1983) have reported adequate screening capabilities (Norris et al., 1987; Okimoto et al., 1982). In elderly medical inpatient screening studies, researchers using the GDS, and a Depression scale from the SCL-90 (Magni et al., 1986) and the BDI and the GDS (Rapp et al., 1988) also reported adequate screening capabilities. The present study extends these findings to a nonmedical, elderly, community-dwelling population.

These results stand in contrast to Toner et al.'s (1988) report of better prediction by interview measures than by paper-and-pencil instruments. Important differences between the samples and the specific instruments used may help account for this discrepancy. Toner et al.'s sample had limited education, and many spoke Spanish as a primary language, whereas our subjects were predominantly middle class and all of them spoke English as their primary language. Toner et al. (1988) noted that many of those who completed the Zung Self-Rating Depression Scale,

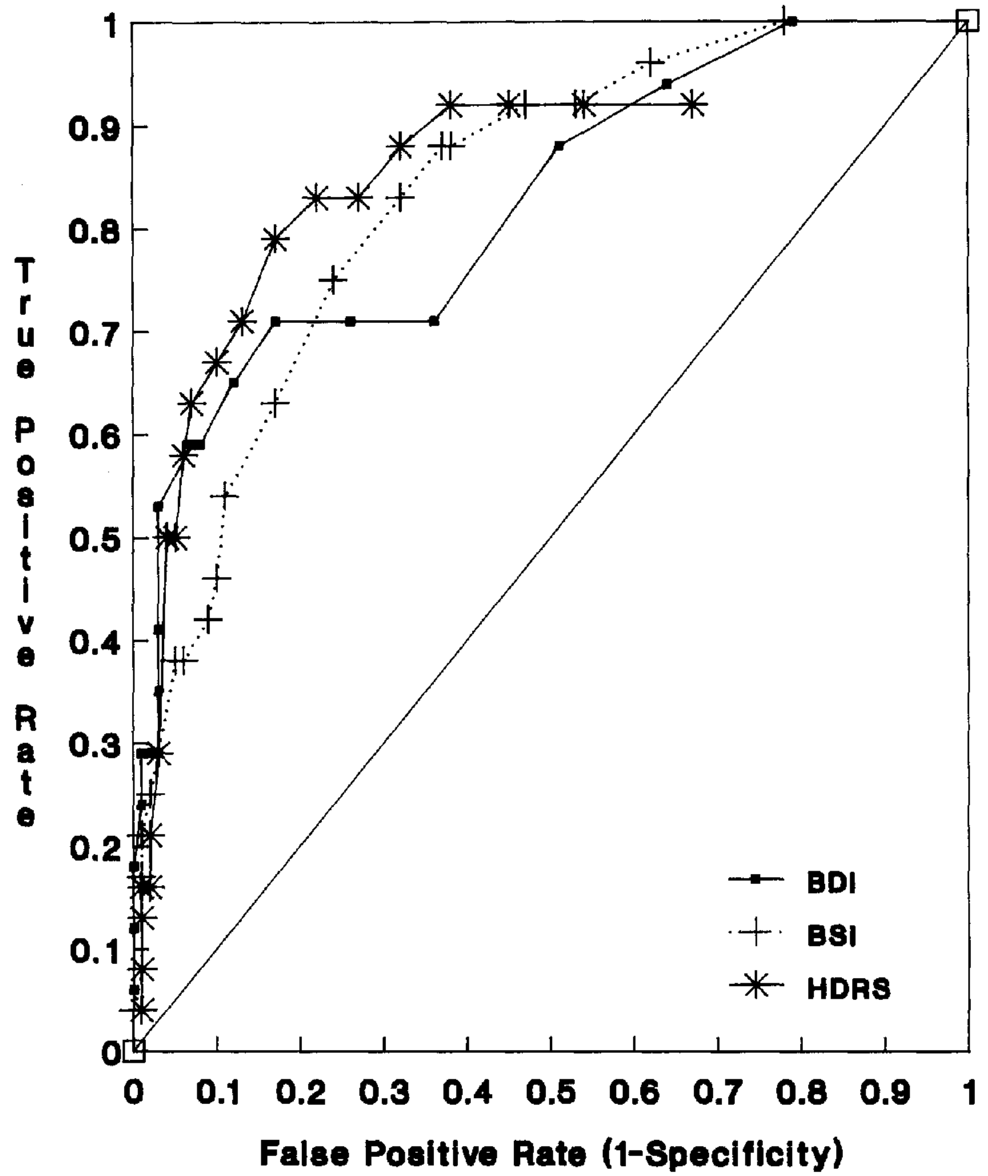

Figure 1. Receiver operating characteristics analysis of comparison: Prediction of clinical depression. (BDI = Beck Depression Inventory; BSI = Brief Symptom Inventory; HDRS = Hamilton Depression Rating Scale)

their self-report instrument, complained of difficulty with reversed items. Neither the BSI Depression scale nor the BDI reversed items.

The sensitivity of all three instruments is higher than the specificity, especially when identifying cases of major depression and depression NOS. This is in keeping with a philosophy that less harm is caused by false positives than by false negatives (Nielsen & Williams, 1980). The BSI Depression scale's sensitivity to men's depression was enhanced by the use of the normal adult norms, as opposed to norms for the elderly (Hale et al., 1984). However, none of the instruments proved to be particularly sensitive to dysthymia in this population. The BDI, with the lowest recommended cutoffs, would have identified fewer than one half of the dysthymia cases. Even the HDRS failed to note any symptomatology in two cases of dysthymia. Because all three instruments focus on symptomatology during the week before the test, whereas dysthymia is determined by symptoms mostly present during the past 2 years, a time factor probably accounts for the screening instruments' relative inability to identify dysthymic cases.

Although there are methodological and item-content differences among the three scales that, with a large enough sample size, should be reflected in the statistical rejection of the null hypothesis (Meehl, 1978), we believe that the current sample size is sufficient to detect those differences worth detecting (see also Meehl, 1978). Indeed, all other things being equal, an infinite sample size would not produce a statistically significant result, because the values of *z* are all less than 1. The incremental utility of the HDRS has, therefore, not been demonstrated with this sample.

In conclusion, we evaluated the relative efficacy of three depression-screening scales in a nonclinical population, and no incremental utility was found for an interview scale over either paper-and-pencil scale. We also demonstrated the applicability of ROC curve analysis to studies with psychological variables. Finally, we replicated prior studies of pencil-and-paper depression screening scales' validity with an older population and pointed to a potential weakness in the scales' ability to identify cases of dysthymia.

References

American Psychiatric Association (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.) Washington, DC: Author.

Beck, A. T., & Beck, R. W. (1972). Screening depressed patients in family practice: A rapid technique. *Postgraduate Medicine, 52,* 81–85.

Blazer, D. G., & Williams, C. D. (1982). Epidemiology of dysphoria and depression in an elderly population. *American Journal of Psychiatry, 137,* 439–444.

Crowne, D. P., & Marlowe, D. (1960). The new scale of social desirability independent of pathology. *Journal of Consulting Psychology, 24,* 349–345.

Derogatis, L. R., & Spencer, P. M. (1982). *The Brief Symptom Inventory (BSI): I. Administration, scoring, and procedures manual.* Baltimore, MD: Johns Hopkins University Press.

Gallagher, D. (1986). The Beck Depression Inventory and older adults. *Clinical Gerontologist, 5,* 149–163.

George, L. K., & Gwyther, L. P. (1986). Caregiver well-being: A multidimensional examination of family caregivers of demented adults. *Gerontologist, 26,* 253–259.

Hale, W. D., Cochran, C. D., & Hedgepeth, B. E. (1984). Norms for the elderly on the Brief Symptom Inventory. *Journal of Consulting and Clinical Psychology, 52,* 321–322.

Hamilton, M. (1967). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23,* 56–62.

Hammen, C. L. (1980). Depression in college students: Beyond the Beck Depression Inventory. *Journal of Consulting and Clinical Psychology, 48,* 126–128.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29–36.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148,* 839–843.

Kiecolt-Glaser, J. K., Dyer, C. S., & Shuttleworth, E. C. (1988). Upsetting social interactions and distress among Alzheimer's disease family caregivers: A replication and extension. *American Journal of Community Psychology, 16,* 825–837.

Koenig, H. G., Meador, K. G., Cohen, H. J., & Blazer, D. G. (1988). Self-rated depression scales and screening for major depression in the older hospitalized patient with medical illness. *Journal of the American Geriatrics Society, 36,* 699–706.

Kozma, A., & Stones, M. J. (1987). Social desirability in measures of subjective well-being: A systematic evaluation. *Journal of Gerontology, 42,* 56–59.

Magni, G., Schifano, F., & de Leo, D. (1986). Assessment of depression in an elderly medical population. *Journal of Affective Disorders, 11,* 121–124.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8,* 283–298.

Murphy, J. M., Berwick, D. M., Weinstien, M. C., Borus, J. F., Budman, S. H., & Klerman, G. L. (1987). Performance of screening and diagnostic tests: Application of receiver operating characteristic analysis. *Archives of General Psychiatry, 44,* 550–555.

Nielsen, A. C., & Williams, T. W. (1980). Depression in ambulatory medical patients: Prevalence by self-report questionnaire and recognition by nonpsychiatric physicians. *Archives of General Psychiatry, 37,* 999.

Norris, J. T., Gallagher, D., Wilson, A., & Winograd, C. H. (1987). Assessment of depression in geriatric medical outpatients: The validity of two screening measures. *Journal of the American Geriatrics Society, 35,* 989–995.

Okimoto, J. T., Barnes, R. F., Veith, R. C., Raskind, M. A., Inui, T. S., & Carter, W. B. (1982). Screening for depression in geriatric medical patients. *American Journal of Psychiatry, 139,* 799–802.

Rapp, S. R., Parisi, S. A., Walsh, D. A., & Wallace, C. E. (1988). Detecting depression in elderly medical inpatients. *Journal of Consulting and Clinical Psychology, 56,* 509–513.

Redick, R. W., & Taube, C. A. (1980). Demography and mental health care of the aged. In J. E. Birren & R. B. Sloane (Eds.), *Handbook of mental health and aging* (pp. 57–71). Englewood Cliffs, NJ: Prentice-Hall.

Riskind, J. H., Beck, A. T., Berchick, R. J., Brown, G., & Steer, R. A. (1987). Reliability of *DSM–III* diagnoses for major depression and generalized anxiety disorder using the structured clinical interview for *DSM–III. Archives of General Psychiatry, 44,* 817–820.

Spitzer, R. L., Williams, J. B. W., Endicott, J., & Gibbon, M. (1987). *Structured Clinical Interview for DSM–III–R—Non-Patient Version (SCID).* New York: New York State Psychiatric Institute.

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlowe-Crowne social desirability scale. *Journal of Clinical Psychology, 28,* 191–193.

Toner, J., Gurland, B., & Teresi, J. (1988). Comparison of self-administered and rater-administered methods of assessing levels of severity of depression in elderly patients. *Journal of Gerontology, 43,* 136–140.

Yeasavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Lehrer, V. O. (1983). Development and validation of a geriatric depression screening scale: A preliminary report. *Journal of Psychiatric Research, 17,* 37–49.

Zemore, R., & Eames, N. (1979). Psychic and somatic symptoms of depression among young adults, institutionalized aged and noninstitutionalized aged. *Journal of Gerontology, 34,* 716–722.

Zung, W. W. K. (1965). A self-rating depression scale. *Archives of General Psychiatry, 41,* 949–958.

Received May 8, 1989
Revision received September 13, 1989
Accepted November 1, 1989 ■