

Comparing Depression Screening Tests in Older Veterans

Stacy Schantz Wilkins, PhD; Jenice Guzman-Clark, RN, PhD; Sheryl Osato, PhD; Nancy Harada, PhD; Donna Henriques, RN, PhD; Catherine Royer, LCSW; Tania Abouezzedine, PhD; Steven Castle, MD; and Shawkat Dhanani, MD

In choosing a depression screening instrument, the practitioner has many factors to consider, including sensitivity and specificity of the test, as well as characteristics of the population being screened.

Depression in older adults not only causes distress, but also can lead to impairments in physical, mental, and social functioning. An estimated 20% of older adults in the community, and as many as 50% in nursing homes, have a diagnosis of depression.¹ The detection and treatment of depressive symptoms in this population is critical to improving quality of life, general health, morbidity, and mortality.

Two of the most widely used depression screening instruments for older adults are the Geriatric Depression Scale (GDS), in use since 1983, and the depression module of the Patient Health Questionnaire (PHQ-9), which became available in 2001.²⁻⁵

Dr. Wilkins is a clinical psychologist and neuropsychologist, **Dr. Guzman-Clark** is a gerontological nurse practitioner, **Dr. Henriques** is a geriatric nurse, **Dr. Abouezzedine** was a psychology intern during the time of this study, **Dr. Castle** is the clinical director, and **Dr. Dhanani** is the director of the Geriatric Evaluation and Management Unit, all at the Geriatric Research, Education and Clinical Center; **Dr. Osato** is the director and **Ms. Royer** is a clinical social worker, both in the Geriatric Psychiatry Outpatient Program; and **Dr. Harada** is the associate director of graduate medical education, all at the VA Greater Los Angeles Healthcare System in California. In addition, Dr. Wilkins is an associate clinical professor of medicine, Dr. Guzman-Clark is an assistant clinical professor of nursing, Dr. Osato is a clinical professor of psychiatry, Dr. Harada is an adjunct professor of medicine, and Drs. Castle and Dhanani are clinical professors of medicine, all at the University of California in Los Angeles.

Both instruments are in the public domain, and thus require no fee to use.

When selecting a depression screening instrument for older adults, particularly those with various ethnic backgrounds and medical and/or psychiatric conditions, the practitioner needs to be aware of each test's sensitivity, specificity, and ease of administration. Therefore, we conducted a comparative study to assess the usefulness of the long and short forms of the PHQ-9, as well as 2 short forms of the GDS, in older, culturally diverse veterans seen in VA geriatric medicine and geriatric psychiatry clinics. We also offer some recommendations on optimal situations for using each of these instruments.

BACKGROUND

The GDS is a 30-item measure that was originally developed for use with older adults to overcome the potential biases that medical illness could introduce into the detection of depression; therefore, it does not include somatic items. Instead, its questions, which use a yes/no format and were developed by geriatric mental health clinicians, include such items as "Do you often feel downhearted and blue?" This format is easier for patients to use than a multiple-category response scale, such as the 5-category Likert scale.

Two short forms of the GDS have been developed; 1 includes 15 questions from the long form, and the other includes 5 questions. These short forms take less time to administer than the 30-item GDS, and their brevity makes the scale easier to use in cognitively impaired individuals. All 3 forms of the GDS were created based on primarily white populations, with relatively small sample sizes.^{2,6,7} The original GDS-30 and its 2 short forms have been shown to be valid and reliable in screening older adults in primary care settings.⁸ In an earlier study of participants in a VA community health screening setting, we found that the short forms of the GDS were reliable indicators of depression for white and Asian groups but not for an African American sample, in whom, the test had poor sensitivity, poor positive predictive power, and low overall diagnostic accuracy.⁹

The PHQ is a self-administered version of the Primary Care Evaluation of Mental Disorders (PRIME-MD®; Pfizer Inc, New York, New York), a 2-step screening and diagnostic instrument for common mental disorders. The PHQ-9 is the depression module; it surveys the patient on how often he/she has been "bothered" during the past 2 weeks by any of the 9 problems the *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, (DSM-IV)*

has identified as diagnostic criteria for major depression. The 4-category scale ranges from 0 (not at all) to 3 (nearly every day). The PHQ-9 was validated in a large multiethnic general medical and obstetric population. The test has good validity and reliability, and its consistency with DSM-IV diagnostic criteria offers an added advantage.^{2,10} The PHQ-9 also provides ratings of depression severity (none, mild, moderate, moderately severe, and severe) based on the respondent's score. The measure has been an effective tool for measuring depression in a variety of settings and patient populations, including older adults.¹¹⁻¹⁴ Additional evidence suggests that the PHQ-9 may have greater sensitivity than more traditional measures, such as the GDS, when assessing depression in culturally diverse populations.¹⁵

The PHQ-2/3 is an abbreviated form of the PHQ-9 that was developed to improve the usefulness of the measure by reducing administration time and the respondent's test-taking burden.⁴ The measure includes the first 2 items of the PHQ-9, which pertain to feelings of depression, hopelessness, and anhedonia. Scores for each item range from 0 to 3, based on symptom frequency, for a possible total of 6. This test has been shown to be a useful tool for depression screening in primary care settings, as well as with older adults.⁵ Current VA practice guidelines call for screening all patients with the PHQ-2/3.¹⁶ The PHQ-2/1, an abbreviated yes/no version of the PHQ-2/3 (rather than the 0- to 3-point scoring for each item), also has been tried, but, so far, has not proved to be useful.¹⁷

In this study, we assessed the utility of various versions of the GDS and PHQ-9—the GDS-15 and GDS-5, and the PHQ-9, PHQ-2/3, and PHQ-2/1—in older, culturally diverse veterans

seen in VA geriatric medicine and geriatric psychiatry clinics. (We included the PHQ-2/1 despite its lack of prior validity, as we have seen many clinicians administer this version.)

Procedure

We asked all new patients, plus some existing patients, without a diagnosis of dementia (dementia diagnosis was determined by medical record prob-

The PHQ-9 may have greater sensitivity than more traditional measures, such as the GDS, when assessing depression in culturally diverse populations.

We evaluated the agreement between the GDS and PHQ-9, based on patient ethnicity and diagnostic group (medical vs psychiatric patients).

METHODS

Participants were screened as part of a clinical improvement project being undertaken in the geriatric psychiatry and geriatric medicine clinics in the VA Greater Los Angeles Healthcare System (VAGLAHS). The clinical demonstration project was developed, in part, to assist in determining which instrument is best to use for our depression screenings. Patients received the GDS (30- or 15-item versions) and the PHQ-9. Subjects in the geriatric medicine clinic were given the 30-item GDS, while those in the geriatric psychiatry clinic were given the shorter GDS-15. In order to evaluate both groups, we extrapolated the GDS-15 from the GDS-30 for the patients in the geriatric medicine clinic by scoring only the GDS-15 items in the 30-item measure. A subset of 40 patients seen in the geriatric medicine clinic also received a clinical diagnostic interview for depression. The Institutional Review Board of the VAGLAHS approved the use of clinical data for research purposes.

lem list), who attended the clinics between 2007 and 2008, to complete the GDS and the PHQ-9. Nursing (medical clinic) or mental health staff (psychiatry clinic) asked patients to complete the forms when they arrived for appointments and were in the waiting room. Tests were self-administered and the data collected were used as a part of routine medical care. The 40 patients in the medical clinic who received a psychiatric interview by a licensed clinical psychologist, in addition to taking the 2 tests, were all new referrals. Diagnostic and demographic data were collected for all participants through medical chart review.

Data analyses

We analyzed study data with statistical package SPSS version 17.0 (SPSS Inc, Chicago, Illinois). Significance was identified with a $P \leq .05$.

We performed 1-way analysis of variance (ANOVA) for continuous variables and chi-square tests (or Fisher's exact test if the numbers in the cells were fewer than 5) for categorical variables to compare demographic characteristics between the 2 clinics and among the 4 cultural groups represented in the sample (African American, Asian, white, and

Hispanic). In addition, we assessed the kappa, or agreement, of the short forms with the long forms of these instruments.^{18,19}

Using results of the clinical interviews as the diagnostic gold standard, we assessed the relationship between test scores and interview findings to determine the sensitivity and specificity of the GDS and PHQ-9 long and short versions. Diagnostic accuracy and sensitivity and specificity rates were manually calculated. We analyzed data with logistic regression to determine the best screening tool to predict the psychologic diagnosis, after controlling for such variables as age and level of education.

By extrapolating from the GDS-30, we determined scores for the 15- and 5-item GDS (with cut-off scores of 5 and 2, respectively). Similarly, we extrapolated scores for the PHQ-2/3 (cut-off score of 3) and PHQ-2/1 (cut-off score of 1) from the PHQ-9, for which the cut-off score is 10. We determined base rates of depression screening by comparing means and SDs for the different groups, as well as percentage of depression prevalence by different tests and cut-off scores.

RESULTS

A total of 155 veterans completed the GDS (30 or 15) and PHQ-9 depression scales during a 1-year period. Mean age of the participants was 78.1 (SD, 7.1) years. We compared patient data to ensure that no veterans were assessed in both clinics.

Compared with patients in the medical clinic, patients in the psychiatry clinic had higher depression scores with greater numbers receiving treatment for depression (Table 1). Almost one-third (31.4%) of the medical patients also were receiving depression treatment, however. Patients in the 2 clinics did not differ

Table 1. Depression treatment rates and screening scores^a

	All (n = 155)	Medical clinic (n = 68)	Psychiatry clinic (n = 87)
Current treatment for depression, %	54.10	31.40	72.40
Screening score, mean (SD)			
GDS-15	5.26 (4.00)	4.75 (4.11)	5.67 (4.09)
GDS-5	1.82 (1.61)	1.69 (1.60)	1.93 (1.62)
PHQ-9	7.81 (7.00)	6.93 (6.42)	8.49 (6.28)
PHQ-2/3	1.94 (1.74)	1.71 (1.77)	2.11 (1.71)

GDS = Geriatric Depression Scale; PHQ = Patient Health Questionnaire.

^aCut-off scores for positive screen for depression: GDS-15 = 5, GDS-5 = 2, PHQ-9 = 10, PHQ-2/3 = 3.

significantly in regard to age or education (Table 2).

When comparing patients based on ethnicity, we found 2 significant differences: First, whites received significantly more education than African Americans (14.75 years vs 12.79 years, respectively). Second, Asians were significantly older (83.2 years) than Hispanics (74.6 years). Given the small number of Asian and Hispanic participants (19 total), we did not conduct further analysis of differences among these cultural groups.

Base rates of positive depression screenings

Scores on the GDS-15 ranged from 0 to 15, with an overall mean of 5.26 (SD, 4.0), with higher scores indicating greater levels of depression. The overall prevalence of a positive depression screening, based on the GDS-15, was 43%. PHQ-9 scores ranged from 0 to 27, with an overall mean of 7.81 (SD, 7.0). Positive depression screening base rate, according to the PHQ-9, was 31%. It is noteworthy that the GDS-15 identified more patients as depressed than the PHQ-9 (66 vs 48, respectively).

ANOVA analysis revealed no significant difference in rate of depres-

sion by clinic or cultural group, though positive depression screening rates were slightly lower in the geriatric medicine group than in the geriatric psychiatry group (39% vs 46% on the GDS-15, and 31% vs 34.5% on the PHQ-9, respectively). Broken down by cultural group, depression was identified in 44% of both the white and the African American patients on the GDS-15, and in 30% of the whites and 32.5% of the African Americans on the PHQ-9.

Instrument agreement

The PHQ-9 and GDS-15 were significantly related ($P \leq .001$ for all, unless noted otherwise) for the total sample, each clinic group, and in both the white and African American groups. The GDS-15 was significantly correlated with the GDS-5, PHQ-9, PHQ-2/3, and PHQ-2/1 ($P = .031$ for whites; $P = .003$ for African Americans).

Test characteristics

Kappa statistics revealed acceptable levels of agreement between the GDS-15 and GDS-5, as well as between the PHQ-9 and PHQ-2/3 for all subjects. No cultural differences were found. The PHQ-2/1 had poor agree-

Table 2. Participants' demographic data

Characteristic	All (n = 155)	Medical clinic (n = 68)	Psychiatry clinic (n = 87)
Age, mean (SD), y	78.1 (7.1)	79.2 (6.6)	77.2 (7.3)
Ethnicity, No. (%)			
White	90 (57.3)	34 (48.6)	56 (64.4)
African American	43 (27.4)	23 (32.9)	20 (23.0)
Asian	10 (6.4)	7 (10.0)	3 (3.4)
Hispanic	9 (5.7)	4 (5.7)	5 (5.7)
Male, %	93.6	88.6	97.7
Education, mean (SD), y	13.9 (3.2) ^a	13.7 (3.0) ^b	14.2 (3.3) ^c

y = years. ^an = 108. ^bn = 55. ^cn = 53.

ment with the PHQ-9 in all assessed groups, however.

The kappas for the PHQ-2/3 and GDS-5 indicate good test agreement with the longer versions. Short forms did identify greater numbers of patients as depressed, however. For example, the GDS-15 identified 66/152 patients (43%) as depressed, while the GDS-5 identified 74/152 patients (49%) as depressed. Furthermore, the PHQ-9 identified 48/155 patients (31%) as depressed, while the PHQ-2/3 identified 61/155 (39%).

GDS and PHQ-9 results and psychiatric interviews

Based on the clinical diagnostic interview, conducted by a licensed clinical psychologist, 8 of the 40 patients in the medical clinic were depressed. Results of the GDS-30 ($P = .014$), GDS-15 ($P = .008$), and PHQ-2/3 ($P = .008$) were all significantly related to the interview findings. The GDS-5 and PHQ-2/1 were not. For this group of 40 geriatric medical patients, the PHQ-9 had a sensitivity of .50 and a specificity of .875. The PHQ-2/3 had a sensitivity of .75 and a specificity of .78. The GDS-30 had

a sensitivity of .75 and a specificity of .75. The GDS-15 had a sensitivity of .75 and a specificity of .78. After controlling for age and education, all screening tools, except the PHQ-9, were found to be significant predictors of a psychiatric interview diagnosis of depression ($P = .057$).

DISCUSSION

The results of this investigation suggest that the GDS and PHQ-9 and their short forms (with the exception of the PHQ-2/1) are valid in white and African American outpatient veterans in both medical and psychiatry clinics. Of note, the GDS identified more subjects as depressed than did the PHQ-9. In addition, the short forms of both instruments had higher rates of depression identification than the longer versions, which supports existing findings in the literature.

Both instruments have been found to be valid for use in older patients, though some studies have questioned the utility of the GDS in certain cultural groups that may be less culturally assimilated than others, such as Chinese Americans, Mexican Ameri-

cans, and certain African American groups.^{17,20-22} Of note, our earlier study found that the short forms of the GDS were valid for white and Asian populations but not for an African American population—a cultural difference that the current study did not replicate.⁹ Our earlier results may have reflected the specific African American group that we screened, which was younger and had less education than the white and African American groups in this study.

Optimal use of the short forms of the GDS and PHQ-9

Compared with the PHQ-9, the GDS identified a higher number of older veterans as depressed, as did the shorter versions of both depression screening scales. Although further study certainly is warranted, clinicians may find this information helpful in choosing the best depression screening scale to fit their population. Clinicians may wish to administer the GDS in situations where it is critical to avoid missing anyone with depression; for example, with patients whose depression may compromise their progress in rehabilitation. Also, it may be useful to balance the length of test administration against the availability of resources to follow up on a false-positive depression screening result.

The VA currently uses the PHQ-2/3 as a screening measure for all veterans because of its brevity; however, the PHQ-2/3 may identify more nondepressed patients than the full PHQ-9 for clinical follow-up (false-positives). In clinics whose patients are able to complete a screening instrument themselves, the full PHQ-9 may be more useful than the shorter version because of potentially fewer false-positive results.

Our study showed that the PHQ-2/1 did not have good sensitivity,

specificity, or test agreement with the full version. The PHQ-2/1 also has been found to have lower sensitivity and specificity than the PHQ-2/3.¹⁶ Therefore, we recommend that practitioners use the PHQ-2/3, not the

ment, rather than a separate gold standard, such as the clinical interview, the ability of the short forms to perform better than the reference scale was limited. The validity of the gold standard and subjects' depression classification

more patients as depressed than the PHQ-9. Based on our results, clinicians may rely on either the GDS (30, 15, or 5), or the PHQ (9 or 2/3) as screening instruments in the clinical care of ethnically diverse geriatric veterans who are receiving care in medical and psychiatry clinics.

We recommend the GDS-15 or GDS-5 for use with older adults presenting in an outpatient psychiatry clinic and for older patients who are medically frail, such as inpatient medical populations, because base rates of depression are higher in these groups, and potentially more critical to identify. These instruments' brevity, along with their yes/no format, make them easy to use.

For older adults who are medically stable and seen in an outpatient setting, we recommend the PHQ-9 (full version). Its 9-item, Likert-based format does not impose a burden on fully functional populations. The full version results in fewer false-positives than the short forms of the instrument, and clinicians can focus their resources on patients with significant levels of depression. ●

Based on our results, clinicians may rely on either the GDS (30, 15, or 5), or the PHQ (9 or 2/3) as screening instruments in the clinical care of ethnically diverse geriatric veterans

PHQ-2/1, for brief screening, as the VA already does for all veterans.

Finally, consideration of the test formats may be helpful. Because the GDS was originally developed to eliminate somatic items, it may be a particularly useful instrument with older patients who have greater levels of physical illness.² The PHQ-9 takes items directly from the *DSM-IV* diagnostic categories and may thereby simplify follow-up clinical interviews, as it can be used as a springboard for symptom identification and management over time.³

Study limitations

Practitioners must consider our results in light of the study's limitations. First, the clinical diagnostic interview—the gold standard for the diagnosis of depression—was conducted in a subgroup of only 40 participants. In addition, while a licensed clinical psychologist conducted this interview, based on *DSM-IV* criteria, all participants, ideally, should have been evaluated with a formal diagnostic measure, such as the Structured Clinical Interview for the *DSM-IV*.²³ By using the PHQ-9 and the GDS-15 as the gold standard when looking at test agree-

ment, rather than a separate gold standard, such as the clinical interview, the ability of the short forms to perform better than the reference scale was limited. The validity of the gold standard and subjects' depression classification

using the GDS and PHQ-9 screening tests is unclear, but the literature associating the GDS and the PHQ-9 with *DSM-IV* diagnostic criteria for major depression is extensive. Another limitation of our study is that our sample comprised outpatient, primarily male, veterans. The extent to which results can be generalized to women, nonveterans, or those in acute or institutional settings remains unclear.

CONCLUSION

Results of this investigation suggest that the GDS-15 and GDS-5, and the PHQ-9 and PHQ-2/3 are valid screening instruments in diverse elderly populations. The PHQ-2/1 was not correlated with results of the psychiatric interview and had low kappas, however.

The PHQ-9, PHQ-2/3, GDS-15, and GDS-5 all were found to be useful screening measures in ethnically diverse patients from both medical and psychiatry clinics. Ethnicity (African American vs white) and diagnosis (medical vs psychiatric) did not play a role in test results. Of note, the GDS and the shorter forms of both the GDS and PHQ did identify

Author disclosures

The authors report no actual or potential conflicts of interest with regard to this article.

Disclaimer

The opinions expressed herein are those of the authors and do not necessarily reflect those of the Federal Practitioner, Quadrant HealthCom Inc., the U.S. Government, or any of its agencies. This article may discuss unlabeled or investigational use of certain drugs. Please review complete prescribing information for specific drugs or drug combinations—including indications, contraindications, warnings, and adverse effects—before administering pharmacologic therapy to patients.

REFERENCES

- American Psychological Association. Public policy, facts about depression in older adults, 2003. <http://www.apa.org/about/gr/issues/aging/mental-health.aspx>. Accessed May 2011.
- Yesavage JA, Brink TL, Rose TL, et al. Development and validation of a geriatric depression screening scale: A preliminary report. *J Psychiatric Res*. 1982;1983;17(1):37-49.
- Kroenke K, Spitzer RL, Williams JB. The PHQ-9: Validity of a brief depression severity measure. *J Gen Intern Med*. 2001;16(9):606-613.
- Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: Validity of a two-item depression screener. *Med Care*. 2003;41(11):1284-1292.
- Li C, Friedman B, Conwell Y, Fiscella K. Validity of the Patient Health Questionnaire 2 (PHQ-2) in identifying major depression in older people. *J Am Geriatr Soc*. 2007;55(4):596-602.
- Sheikh JI, Yesavage JA. Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. *Clin Gerontol*. 1986;5(1-2):165-173.
- Hoyl MT, Alessi CA, Harker JO, et al. Development and testing of a five-item version of the Geriatric Depression Scale. *J Am Geriatr Soc*. 1999;47(7):873-878.
- Lyness JM, Noel TK, Cox C, King DA, Conwell Y, Caine ED. Screening for depression in elderly primary care patients. A comparison of the Center for Epidemiologic Studies-Depression Scale and the Geriatric Depression Scale. *Arch Intern Med*. 1997;157(4):449-454.
- Barclay TR, Wilkins SS, Harada ND, Osato S, Dhanani S. Detecting depression in older veterans: Impact of cultural diversity on test validity. *Fed Pract*. 2008;25(4):28-33.
- American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, 4th ed, Text Revision*. Washington, DC: American Psychiatric Association; 2000.
- Ell K, Unützer J, Aranda M, Sanchez K, Lee PJ. Routine PHQ-9 depression screening in home health care: Depression, prevalence, clinical and treatment characteristics and screening implementation. *Home Health Care Serv Q*. 2005;24(4):1-19.
- Lee PW, Schulberg HC, Raue PJ, Kroenke K. Concordance between the PHQ-9 and the HSCL-20 in depressed primary care patients. *J Affect Disord*. 2007;99(1-3):139-145.
- Martin A, Rief W, Klaiberg A, Braehler E. Validity of the brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. *Gen Hosp Psychiatry*. 2006;28(1):71-77.
- Williams LS, Brizendine EJ, Plue L, et al. Performance of the PHQ-9 as a screening tool for depression after stroke. *Stroke*. 2005;36(3):635-638.
- Chen TM, Huang FY, Chang C, Chung H. Using the PHQ-9 for depression screening and treatment monitoring for Chinese Americans in primary care. *Psychiatr Ser*. 2006;57(7):976-981.
- Office of Quality and Performance (10Q). FY 2008, Q1 technical manual for the VHA performance measurement system. Washington, DC: Veterans Health Administration; 2007.
- Corson K, Gerrity MS, Dobscha SK. Screening for depression and suicidality in a VA primary care setting: 2 items are better than 1 item. *Am J Manag Care*. 2004;10(11 Pt 2):839-845.
- Blackman NJ, Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. *Stat Med*. 2000;19(5):723-741.
- Baker FM, Espino DV, Robinson BH. Assessing depressive symptoms in African American and Mexican American elders. *Clin Gerontol*. 1993;14(1):15-29.
- Flacker JM, Spiro L. Does question comprehension limit the utility of the Geriatric Depression Scale in older African Americans? *J Am Geriatr Soc*. 2003;51(10):1511-1512.
- Mui A. Geriatric Depression Scale as a community screening instrument for elderly Chinese immigrants. *Int Psychogeriatr*. 1996;8(3):445-458.
- Mui AC, Kang SY, Chen LM, Domanski MD. Reliability of the Geriatric Depression Scale for use among elderly Asian immigrants in the USA. *Int Psychogeriatr*. 2003;15(3):253-271.
- First MB, Gibbon M, Spitzer RL, Williams JBW, Benjamin LS. *Structured Clinical Interview for DSM-IV Axis I Personality Disorders, (SCID-I/P)*. Washington, DC: American Psychiatric Press, Inc; 1997.

THE HEALTH PROFILE OF THE DEPARTMENT OF DEFENSE

View online at www.fedprac.com

Developed collaboratively by
Federal Practitioner and Pfizer Inc.

NPC03628B/282538-01 © 2011 Pfizer Inc. and Quadrant HealthCom Inc.
All rights reserved. Printed in the USA/May 2011.

A Supplement to *Federal Practitioner*® Vol. 28 Sup

THE HEALTH PROFILE OF THE DEPARTMENT OF DEFENSE

ACTIVE MILITARY DATA ON:

- ★ Cardiovascular Health
- ★ Obesity
- ★ Mental Health
- ★ Substance Use
- ★ Pain Management
- ★ HEDIS Measures

FEDERAL PRACTITIONER Pfizer