#6

ORIGINAL CONTRIBUTION

Improving the Efficiency of Screening for Major Depression in People With Spinal Cord Injury

Daniel E. Graves, PhD[1]; Charles H. Bombardier, PhD[2]

[1]Department of Physical Medicine and Rehabilitation, Baylor College of Medicine, Houston, Texas; [2]Department of Rehabilitation Medicine, University of Washington School of Medicine, Seattle, Washington

Received April 9, 2007; accepted October 22, 2007

Abstract

Background/Objective: To investigate the metric properties, relative efficiency, sensitivity and specificity, and positive predictive value of a short form of the Patient Health Questionnaire-9 (PHQ-9) that may be used as a screening test for depression.

Methods: Data from the National Spinal Cord Injury Statistical Center Database containing 3,652 records with complete data for the PHQ-9 were analyzed using Confirmatory Factor Analysis, Item Response Theory Graded Response Model analysis, and sensitivity and specificity analysis of classification.

Results: A scale comprised of items 1, 2, and 6 from the PHQ-9 has a relative efficiency of 0.66 compared to the 9-item scale. Using this 3-item scale and a cutoff score of 3 or more provides specificity of 0.93 and sensitivity of 0.87; a cutoff of 4 provides specificity of 0.95 and sensitivity of 0.82. The shorter version of the scale reduces the effect of response bias caused by gender. The relative efficiency of the 9-item scale is 0.88 for women compared to men; the 3-item scale increases the relative efficiency to 0.93.

Conclusion: A 3-item scale provides adequate information for clinical screening purposes. Cutoff scores of either 3 or 4 are acceptable and present options for decision making within a particular clinical setting. Additionally, the 3-item scale reduces the effect of gender of the respondent on the score obtained.

J Spinal Cord Med. 2008;31:177–184

Key Words: Spinal cord injuries; Measurement; Depression, major; Screening

INTRODUCTION

Major depression (MD) is a highly prevalent and disabling secondary condition associated with spinal cord injury (SCI) (1–4). The prevalence of probable MD after SCI ranges from 9.8% (5) to 35% (6) among inpatients and from 13% (7) to 31% (8) among people residing in the community. Depression is associated with longer lengths of hospital stay and fewer functional improvements (9) as well as less functional independence and poorer mobility at discharge (10). Depression is associated with the occurrence of pressure sores and urinary tract infections (11), poorer self-appraised health (12), less leisure activity (13), poorer community mobility and social integration, and fewer meaningful social pursuits (7,8). Persons with SCI and significant depression spend more days in bed and fewer days outside the home, require greater use of paid personal care, and incur greater medical expenses (14). Moreover, symptoms consistent with depression, such as documented expressions of despondency, hopelessness, shame, and apathy, are the variables most predictive of suicide 1 to 9 years after SCI (15).

Given the prevalence and negative impact of MD in this population, early identification and treatment have the potential to reduce unnecessary suffering and impairment in people with SCI. In primary care settings systematic screening for depression plus feedback to providers has been found to improve recognition of MD two- to three-fold (16). As a result the US Preventative Services Task Force has concluded that screening for MD in primary care settings improves outcomes, especially when screening is linked to systematic treatment and monitoring. Our ultimate goal is to affect clinical care for MD among people with SCI in a similar manner to improve detection, treatment, and outcomes.

Please address correspondence to Daniel E. Graves, PhD, Baylor College of Medicine, 1333 Moursund, A-222, Houston, TX 77030; phone: 713.799.5023; fax: 713.799.5030 (e-mail: dgraves@bcm.tmc.edu).

This work was supported by funds from grants H133N000004, H133N060003, H133N000005, H133N000003, and H133N060033 from the National Institute on Disability and Rehabilitation Research in the Office of Special Education and Rehabilitation Services in the US Department of Education.

Table 1. Patient Health Questionnaire-9 Items

Over the last 2 weeks, how often have you been bothered by any of the following problems? Response options: (0) not at all, (1) several days, (2) more than half the days, (3) nearly every day.
1. Little interest or pleasure in doing things.
2. Feeling down, depressed, or hopeless.
3. Trouble falling or staying asleep, or sleeping too much.
4. Feeling tired or having little energy.
5. Poor appetite or overeating.
6. Feeling bad about yourself—or that you are a failure or have let yourself or your family down.
7. Trouble concentrating on things, such as reading the newspaper or watching television.
8. Moving or speaking so slowly that other people could have noticed? Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual.
9. Thoughts that you would be better off dead or of hurting yourself in some way.
If the subject endorses any symptoms of depression ask, "How difficult have these symptoms made it for you to do your work, take care of things at home, or get along with other people?" (1) Not difficult at all, (2) Somewhat difficult, (3) Very difficult, (4) Extremely difficult.

Early detection hinges on having MD screening measures that are reliable, valid, and feasible to implement in a variety of clinical settings. Due to a variety of factors, depression screening measures are not equally effective in medical populations (17). Moreover, the concomitant neurological and medical symptoms associated with SCI likely require validity studies or instrument modifications to improve detection and reduce false-positive results. Therefore, we have begun to investigate the clinical utility of the Patient Health Questionnaire-9 (PHQ-9) (Table 1), one of the most promising depression screening measures (18,19). The PHQ-9 is only 9 items, exactly parallels Diagnostic and Statistical Manual (DSM) IV MD criteria, and has been shown to have good diagnostic accuracy and validity among medical patients (20). Preliminary data on people with SCI show that the PHQ-9 demonstrates good internal consistency, construct validity, and item-level predictive validity (1).

Although the PHQ-9 is relatively brief, even more abbreviated screening measures are desirable. Shorter measures have greater potential to be included in comprehensive health screening batteries or a routine review of systems. The use of fewer items reduces respondent burden and makes it more practical to conduct the multifaceted screening advocated in rehabilitation settings. Prior research indicates that as few as 2 questions can serve as an effective screen for depression (21,22). In a Veterans Affairs urgent care center, 2 questions: "During the past month have you felt down, depressed, or hopeless?" and "During the past month have you been bothered by little interest or pleasure in doing things?" together were nearly as sensitive and specific for diagnosing MD as several much longer measures (22). The first 2 items of the PHQ-9 have also been found to have adequate construct and criterion validity. A score of greater than or equal to 2 on the PHQ-2 had a sensitivity of 83% and a specificity of 92% compared to a structured diagnostic assessment among 580 primary care patients (21).

In this study, our overall goal was to determine the most efficient way to identify persons with probable MD among people with SCI based on the PHQ-9. In contrast with prior studies that used a priori item selection procedures and traditional statistical approaches, we proposed to use Item Response Theory to determine: (a) the most discriminating items; (b) the relative efficiency of scales of varying length compared to the 9-item scale; (c) the most efficient cutoff score for screening; (d) whether basic demographic factors such as gender, time since injury, education level, and race/ethnicity were associated with differences in item response patterns.

MATERIALS AND METHODS

Participants

Participant data were obtained from 16 Model SCI System centers located throughout the United States, including both urban and rural catchment areas. The study sample consisted of 3,652 persons with traumatic SCI who participated in the National Spinal Cord Injury Statistical Center Database from October 2000 through April 2003. All participants were at least 18 years old and provided informed consent for data collection. Each Model System obtained approval from its local institutional review boards for the study and provided a mechanism for clinically managing subjects identified as having serious suicidal ideation.

Probable Major Depression. As described earlier, the PHQ-9 was developed to facilitate identification and diagnosis of DSM IV MD in medical samples (19). To be consistent with DSM IV MD diagnostic criteria, each of the 9 items is rated according to how persistent the symptom has been over the past 2 weeks: 0 (not at all), 1 (several days), 2 (more than half the days), or 3 (nearly every day) (Table 1). A diagnosis of MD can be approximated by comparing item responses directly to the DSM IV diagnostic criteria. However, a large validity study found that the best approach to identifying those with MD in a primary care setting was obtained by calculating a total PHQ-9 score and using a cutoff of greater than or equal to 10 (18). Compared to an independent structured diagnostic interview for MD conducted by a mental health professional, a score of at least 10 on the PHQ-9 had a sensitivity of 88% and a specificity of 88%. In addition, construct validity was documented through significant correlations between

increasing levels of depressive symptoms and poorer health-related quality of life, greater disability days, and more physician visits.

Statistical Analyses

Determining whether the PHQ-9 is a unidimensional measure of the depression construct was a necessary first step in this investigation. A confirmatory factor analysis was conducted to determine whether the items of the PHQ-9 have a unidimensional structure using AMOS 6.0. This was followed by an analysis of the PHQ-9 data using the Graded Response Model, one of several 2-parameter Item Response Theory models that can be applied to polytomous response categories (23,24), utilizing Multilog v7.03 (25). Additionally, each participant was classified as nondepressed vs having probable MD based on whether he/she obtained a total score of less than 10 vs 10 or more on the PHQ-9, consistent with the validity study conducted by Kroenke and colleagues (18). Additionally, the sensitivity, specificity, and positive predictive value of the resultant scale(s) was determined using the Kroenke definition of probable MD as the criterion.

Relative Efficiency

The Graded Response Model analysis provides item category threshold estimates called difficulty estimates and an item discrimination estimate for each item. Unlike classical test theory, Item Response Theory allows the standard error, an estimate of the accuracy of measurement, to vary over a range of ability levels (theta θ). There is no natural or inherent scale for the theta estimates. For ease of interpretation these estimates are centered on zero with a standard deviation of 1. The summary statistic of interest for this investigation is the test information function. Information is inversely related to the standard error. If the standard error is small the amount of information provided is high; likewise, a large standard error will produce low levels of information (26). The information functions for tests of varying length or applied to different populations can be compared using a concept called relative efficiency (27). This is possible because the test information function is the summation of the item information functions that are determined by the item parameters of each item. The items that best discriminate will provide more information relative to the other items in the test.

One property of a good test is that the threshold and discrimination parameters will be invariant between groups. That means that the items should work on all groups equally. If the item parameters are substantially different between groups this is called differential item function (28). Differential item function can affect the validity of a test application if it produces significantly different results for groups. The overall effects of differential item function can be assessed using relative efficiency, also. Since the information function will summarize the accuracy of measurement according to all item parameters, the cumulative effect of differences in parameters can be assessed even if the differences between individual parameters from the same item applied to different groups are not large enough to reach statistical significance.

As with all measures, it is assumed that the items on the PHQ-9 are related to a single underlying dimension. The dimension of interest with the PHQ-9 is MD. This dimension will range from very low scores (less than 0 on the theta scale), meaning no depressive symptoms, to very high scores, meaning depressive symptoms are present. Since the PHQ-9 is intended to provide a screen for depression, it is important that the scale provide the greatest amount of information at the point along this dimension that maximally separates persons with MD from those without MD. These 2 important issues will be addressed separately. First, the confirmatory factor analysis will test the 9 items for fit to a single dimension. Subsequently, the Graded Response Model analysis will develop the information function for any alternative scale configurations.

RESULTS

The sample of 3,652 participants comprised 2,863 male and 789 female patients. The age at time of injury ranged from 4 to 88 years with a mean of 31.8 (13.62) years. The interviews were conducted between 1 and 30 years post injury, with a mean of 9.6 (8.19) years. The average age at time of interview was 41.4 (13.44) years with a range of 18 to 90 years of age. The sample was 77% white, 17% African American, and 6% other or unknown.

Dimensionality

The confirmatory factors analysis demonstrates that the 9 items contained in the PHQ-9 do form a single dimension. The fit statistics for this model show that it is a good fit, but not a perfect fit. The several fit indices provide somewhat different information about the nature of the fit of this model. An index of global fit provided by the chi-square statistic ($\chi^2 = 847$, $df = 27$, $P < 0.001$) indicates that there is still some improvement possible in this model. The goodness of fit index indicates that 95% of the variance is accounted for with the unidimensional structure shown in Figure 1. The Root Mean Square Error of Approximation, however, shows this model to be a fair fit. The value of 0.091 is less than the rule-of-thumb limit for adequate fit of 0.10. The entire 90% confidence interval of the Root Mean Square Error of Approximation, extending from 0.086 to 0.097, is less than the value of 0.10, suggesting an adequate fit. The Normed Fit Index is an index of the fit of this model in relation to 2 theoretical models, the independence model, one with no interrelations at all, and a saturated model, one with all possible interrelations. This value shows that the model is a 92.4% better fit than the independence model. These values

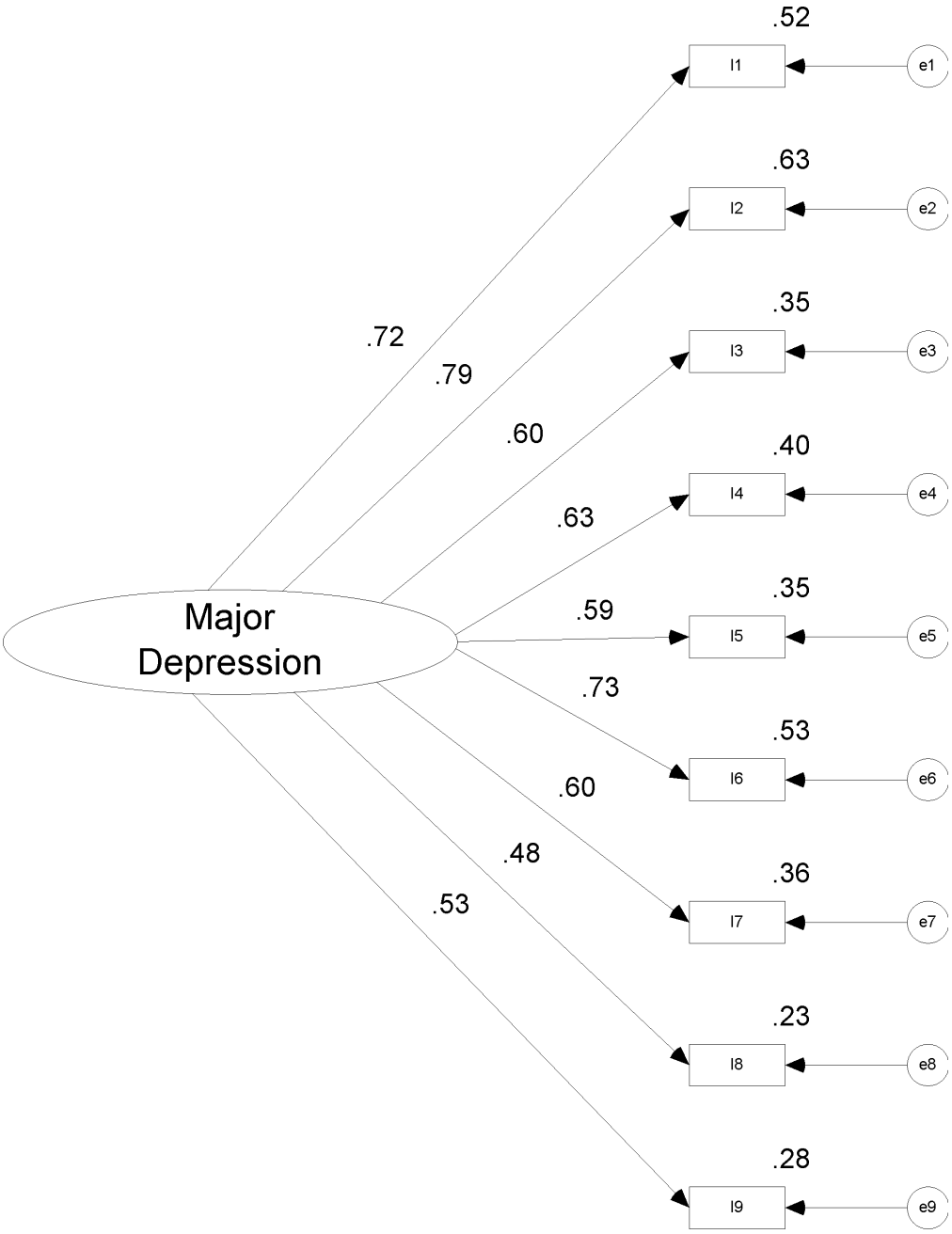

Figure 1. *Confirmatory factor analysis showing the relation of each of the 9 items of the PHQ-9 to the underlying construct.*

indicate that all items of the PHQ-9 do adequately fit to a single underlying construct.

Other values from Figure 1 provide insight into how these items work. The number on the lines between the latent variable and the variable is a correlation coefficient. These values show that items 1, 2, and 6 are the most strongly related to the underlying construct. For all of these variables the correlation between the item and the construct is over 0.70. This means that more than 50% of the variance of these items is related to the construct underlying these items (depression). This is important in that items 1 and 2 are the cardinal signs of depression, depressed mood and anhedonia, while item 6 is a secondary diagnostic symptom having to do with feeling bad about oneself or a sense of failure. A scale with the first 2 items has been utilized previously. The strength of the relation between item 6 and the latent variable indicates that this item may contribute valuable information as well. None of the other items is as strongly related to depression; thus, the others will only be included in the 9-item scale. Therefore, 3 scales will be considered in the following analyses, these are: (a) the original 9 items; (b) items 1 and 2, and (c) items 1, 2, and 6.

Relative Efficiency

The 9-item, 2-item, and 3-item scales for the entire sample of 3,652 records were separately analyzed using the Graded Response Model. Subsequently, the data were analyzed separately for groups based on gender, time since injury, and ethnicity. Figure 2 shows the

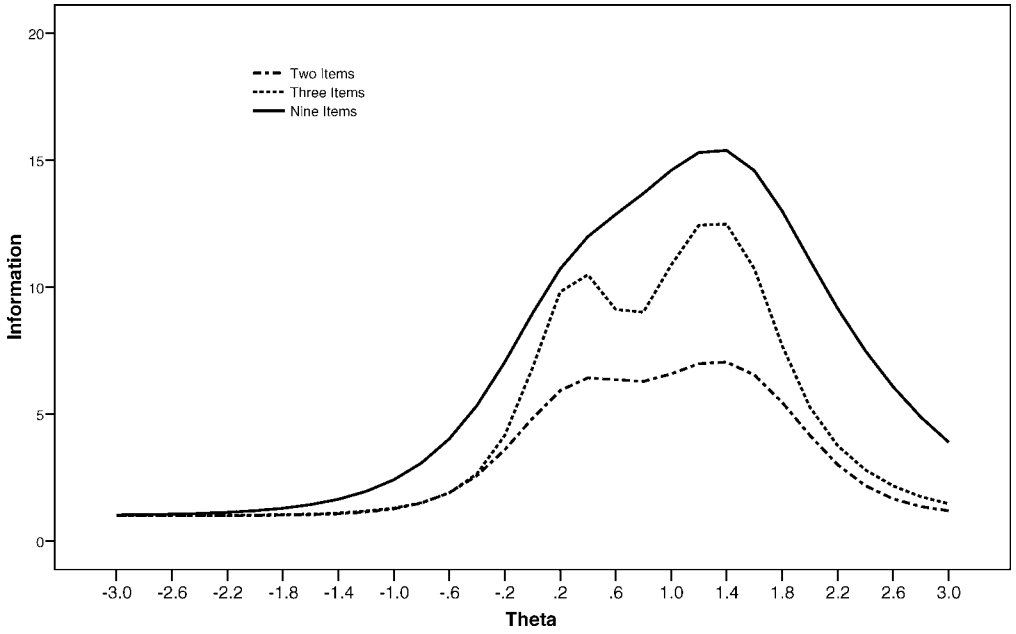

Figure 2. *Test information functions for the 2-, 3-, and 9-item scales.*

information functions for the 3 scale configurations. Relative efficiency of a test is an index of the amount of information available when the length of the test is altered. In this case the relative efficiency will represent the proportion of information available in the shorter scales relative to the information available in the 9-item scale. The relative efficiency of the 2-item test is 0.46. That means that less than half of the information of the 9-item test is available in the 2-item test. However, the 3-item test captures more information, with a relative efficiency index of 0.67. That means that two thirds of the information is available with one third of the items. Figure 2 also shows that all 3-scale configurations seem to reach the peak of information at about the same location (approximately 1.4) on the theta scale. This indicates that the 3 scales provide the most precise measurement in the same range of test scores.

Figure 3 shows the information functions for the entire sample compared to the information functions for the male and female subsamples. The relative efficiency of the 9-item test for men is 1.05 compared to the entire sample, while it is only 0.88 for women. Comparing the relative efficiency for the women vs the men yields 0.83. There are distinct qualitative differences in the discrimination and difficulty parameters when comparing male and female samples; however, none of these differences in individual item parameters is substantial enough to reach statistical significance when tested. The information function would demonstrate the cumulative effect of these differences.

Figure 4 shows the information function for the 3-item scale for all participants compared to the male and female subsamples. The cumulative effect of differential item functioning between male and female subsamples is less with the 3-item scale. The relative efficiency of the 3-item scale for women compared to the entire sample is 0.93; whereas it is 1.02 for men. Comparing the 3-item scale for women to men produces a relative efficiency of 0.90.

The 9-item scale does not demonstrate differential item function in other group comparisons. Specifically, the relative efficiency of the 9-item scale for whites compared to nonwhites is 0.97; time since injury 2 years or less compared to more than 2 years is 1.01; education 12 years or less compared to more than 12 years, 0.96.

Sensitivity and Specificity

The squared correlation coefficient between the total scores on the 3-item scale and the 9-item scale is 0.794. This means that the 3-item score accounts for approximately 79% of the variance in the 9-item total score. Using a cutoff score of 4 on the 3-item scale, 16.5% of the sample was identified as having probable depressive disorder. Using an alternative cutoff of a total score of 3, 25.8% of the sample was identified as having probable depressive disorder. Using a total score of 10 or more on the 9-item test as a criterion, 566 participants were designated as having probable depressive disorder, compared to 941 and 603 so designated using cutoff scores of 3 or 4 on the 3-item scale, respectively.

The sensitivity and specificity of the 3-item scale were calculated using both the total score of 3 and total score of 4 with the total score of 10 or more on the 9-item scale as the criterion. The results shown in Table 2 indicate that the cutoff of a total score of 3 may increase the rate of false positives. An additional 10.4% of the sample is designated using this criterion compared with the total score of 10 points on the 9-item scale. However, the specificity and sensitivity are both acceptable. The cutoff

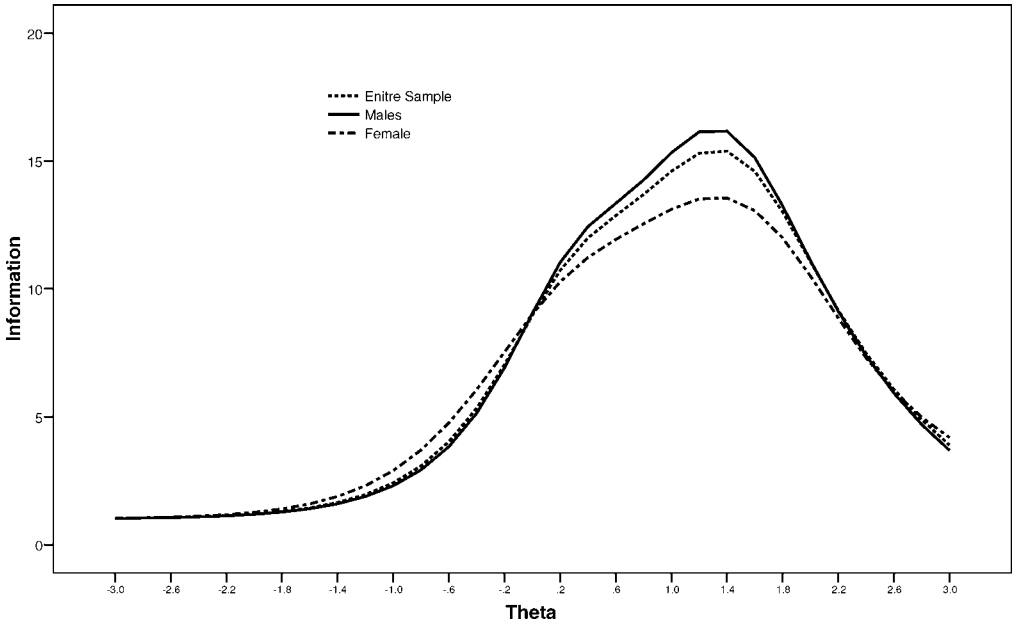

Figure 3. *Information functions demonstrating the difference between male and female subsamples relative to the entire sample for the 9-item scale.*

of a total score of 4 identifies only 1.1% more cases as having probable MD than the criterion of a total score of 10 or more points. The total score of 4 points also has acceptable specificity and sensitivity ratings. As expected, the specificity of the total score of 4 criterion is somewhat higher than the specificity of the total score of 3. This indicates a reduction in the number of possible false positives with the higher cutoff value. The number of possible false positives affects the positive predictive value for each cutoff score. The cutoff of 3 has a positive predictive value of 0.56; the score of 4 has a positive predictive value of 0.77.

DISCUSSION

There is a desire for a shortened depression screening scale that can accurately identify persons with probable MD in need of treatment. The results of this investigation show that a 3-item scale containing items (1) "Bothered by little interest or pleasure in doing things?" (2) "Bothered by feeling down, depressed, or hopeless?" and (6) "Bothered by feeling bad about yourself—or that

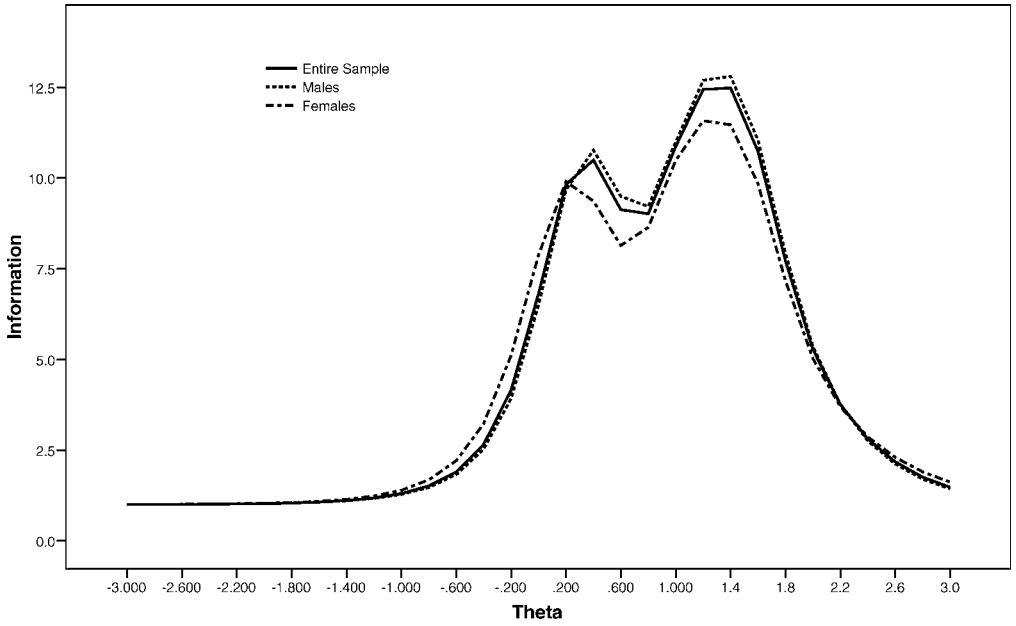

Figure 4. *Information functions for the male and female subsamples relative to the entire sample for the 3-item scale.*

Table 2. Sensitivity, Specificity, and Positive Predictive Value for the 3-Item Screening Test With a Total Score Cutoff of 3 and 4

Score Cutoff		Kroenke Criteria +	Kroenke Criteria −	Kroenke Criteria Total	Specificity	Sensitivity	Positive Predictive Value
3	+	528	413	**941**	0.93	0.87	0.56
	−	38	2,673	**2,711**			
	Total	**566**	**3,086**	**3,652**			
	+	462	141	**603**			
4	−	104	2,945	**3,049**	0.95	0.82	0.77
	Total	**566**	**3,086**	**3,652**			

you are a failure or have let yourself or your family down?" from the PHQ-9 demonstrate the best measurement properties relative to the other scale configurations. These 3 items have the best discrimination capacity compared to the 9 original items. This means that they are not only the most strongly related to the underlying construct of depression, but that they provide the most accurate distinction between levels of depression. This scale structure, in addition to the utility of the cutoff scores described above, seems to be a valid short form of the PHQ-9 for use among people with SCI.

Previous research has shown that the first 2 items of the PHQ-9 have acceptable screening properties (21). The results of this investigation show that the addition of a single item having to do with feeling bad about oneself or like a failure significantly improves the measurement properties of the 2-item scale in this population. The magnitude of improvement in measurement accuracy associated with the addition of this 1 item, an increase from less than half of the information in the PHQ-9 to capturing over two thirds of the information in the total scale, clearly justifies the addition of this item.

This investigation provides options for the clinician. Having classification information for both cutoff scores of 3 and 4 allows for tailoring the administration of the 3-item scale to the need of the clinical setting. In some cases, a lower cutoff and higher sensitivity are preferred if minimizing the false-negative rate is especially important. On the other hand, if the site needs to maximize specificity and minimize the costs associated with follow-up examinations of those who screen positive, the higher cutoff may be more appropriate. Clinicians in each setting must weigh the pros and cons of the alternative cutoff scores to determine the best fit for their circumstances.

Another advantage with the 3-item version of the PHQ described here is how it functions similarly among men and women. The original PHQ-9 item parameters contain many nonsignificant differences when calibrated on male and female samples separately. These differences are not large enough to be statistically significant for any single item in isolation. However, the cumulative effect of these small differences results in significant variation in how the measure operates among men and women once the total score is considered. The 3-item scale reduces the disparity of the response differences between men and women.

The scores on many measures of depression can be influenced by the presence of somatic symptoms that may be etiologically quite ambiguous in people with comorbid medical illness (29). There is controversy regarding the diagnostic utility of somatic symptoms in people with SCI, with some authors indicating that somatic symptoms should not be ignored (1,30), and some emphasizing that psychological symptoms are more important (5). The 3 items found to be most discriminating in this research are all psychological in nature. Therefore, this version avoids the controversy surrounding inclusion of somatic items and should be acceptable to clinicians regardless of their opinion on this issue. In addition, these same 3 items represent the most discriminating symptoms of MD found in other studies of depresssion in people with medical comorbidities (29). If the criterion of either 3 or 4 points is reached using the 3-item scale, this should not be taken as a diagnosis of depression, only an indication that further investigation is needed. The 3-item scale is intended to offer a short screen for the most discriminating symptoms of depression. Further study is required to determine if the use of the 3-item scale functions as well in an inpatient setting as it does in the outpatients participating in this study.

Finally, the only criterion available for this investigation was the longer version of the same scale. This is not optimal for testing the accuracy of a diagnostic instrument. Further studies are required to determine whether the 3-item scale operates as effectively when other measures or clinical interviews are utilized as a criterion. Until such studies are conducted, the results of the 3-item scale should only be used as a first step in the diagnosis of depression.

CONCLUSIONS

Items 1, 2, and 6 from the PHQ-9 provide 66% of the information that can be obtained from the 9-item scale. Cutoff scores of 3 or 4 points for this 3-item version are both acceptable and present reasonable options for

decision making within a particular clinical setting. Additionally, the 3-item scale reduces the effect of the gender of the respondent on the score obtained. Shorter measures like this may fit well into busy clinical practices and help boost rates of screening for MD in this population. One efficient screening approach would involve clinicians asking these 3 questions first. Then, if the patient scores above the chosen cutoff, the remainder of the PHQ-9 could be administered immediately to improve the diagnostic specificity of the assessment.

REFERENCES

1. Bombardier CH, Richards JS, Krause JS, Tulsky D, Tate DG. Symptoms of major depression in people with spinal cord injury: implications for screening. *Arch Phys Med Rehabil.* 2004;85:1749–1756.
2. Elliott TR, Frank RG. Depression following spinal cord injury. *Arch Phys Med Rehabil.* 1996;77:816–823.
3. Krause JS, Kemp B, Coker J. Depression after spinal cord injury: relation to gender, ethnicity, aging, and socioeconomic indicators. *Arch Phys Med Rehabil.* 2000;81:1099–1109.
4. Tate DG, Forchheimer M, Maynard F, Davidoff G, Dijkers M. Comparing two measures of depression in spinal cord injury. *Rehabil Psychol.* 1994;38:53–61.
5. Frank RG, Chaney JM, Clay DL, et al. Dysphoria: a major symptom factor in persons with disability or chronic illness. *Psychiatry Res.* 1992;43:231–241.
6. Kennedy P, Rogers BA. Anxiety and depression after spinal cord injury: a longitudinal analysis. *Arch Phys Med Rehabil.* 2000;81:932–937.
7. MacDonald MR, Nielson W, Cameron M. Depression and activity patterns of spinal cord injury persons living in the community. *Arch Phys Med Rehabil.* 1987;68:339–343.
8. Fuhrer MJ, Rintala DH, Hart KA, Clearman R, Young ME. Depressive symptomatology in persons with spinal cord injury who reside in the community. *Arch Phys Med Rehabil.* 1993;74:255–260.
9. Malec J, Neimeyer R. Psychologic prediction of duration of inpatient spinal cord injury rehabilitation performance of self care. *Arch Phys Med Rehabil.* 1983;64:359–363.
10. Umlauf R, Frank RG. A cluster-analytic description of patient subgroups in the rehabilitation setting. *Rehabil Psychol.* 1983;28:157–167.
11. Herrick S, Elliott T, Crow F. Social support and the prediction of health complications among persons with SCI. *Rehabil Psychol.* 1994;39:250.
12. Schulz R, Decker S. Long-term adjustment to physical disability: the role of social support, perceived control, and self-blame. *J Pers Soc Psychol.* 1985;48:1162–1172.
13. Elliott T, Shewchuck R. Social support and leisure activities following severe physical disability: testing the mediating effects of depression. *Basic Appl Soc Psychol.* 1995;16:471–487.
14. Tate D, Forchheimer M, Maynard F, Dijkers M. Predicting depression and psychological distress in persons with spinal cord injury based on indicators of handicap. *Am J Phys Med Rehabil.* 1994;73:175–183.
15. Charlifue S, Gerhart K. Behavioral and demographic predictors of suicide after spinal cord injury. *Arch Phys Med Rehabil.* 1991;72:488–492.
16. Pignone MP, Gaynes BN, Rushton JL, Burchell CM, Orleans CT, Mulrow CD. Screening for depression in adults: a summary of the evidence for the US Preventive Services Task Force. *Ann Intern Med.* 2002;136:765–776.
17. Williams JW, Noel PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? *JAMA.* 2002;287:1160–1170.
18. Kroenke K, Spitzer R, Williams J. The PHQ-9: validity of a brief depression symptom severity measure. *J Gen Intern Med.* 2001;16:606–613.
19. Spitzer R, Kroenke K, Williams J. Validity and utility of a self-report version of PRIME-MD: The PHQ Primary Care Study. *JAMA.* 1999;282:1737–1744.
20. Kroenke K, Spitzer RL, Williams JBW. Validity of a brief depression severity measure. *J Gen Intern Med.* 2001;16:606–613.
21. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Med Care.* 2003;41:1284–1292.
22. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression. Two questions are as good as many. *J Gen Intern Med.* 1997;12:439–445.
23. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometr Monogr.* 1969;17.
24. Samejima F. Graded response model. In: van der Linden WJ, Hamilton RK, eds. *Handbook of Modern Item Response Theory.* New York, NY: Springer-Verlag; 1997:85–100.
25. Thissen D. *Multilog User's Guide: Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory*. 1991. Version 6.0 ed. Chicago, IL: Scientific Software, Inc; 1991.
26. Lord FM. Information functions and optimal scoring weights. In: Lord FM, ed. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc; 1980:65–82.
27. Lord FM. The relative efficiency of two tests. In: Lord FM, ed. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc; 1980:83–105.
28. Holland PW, Wainer H. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993.
29. Kroenke K. A 75-year-old man with depression. *JAMA.* 2002;287:1568–1576.
30. Clay D, Hagglund K, Frank R, Elliott T, Chaney J. Enhancing the accuracy of depression diagnosis in patients with spinal cord injury using Bayesian Analysis. *Rehabil Psychol.* 1995;40:171–180.