Rasch Analysis of the Geriatric Depression Scale–Short Form

Karl S. Chiang, MA, Kathy E. Green, PhD, Enid O. Cox, PhD
Purpose: The purpose of this study was to examine scale dimensionality, reliability, invariance, targeting, continuity, cutoff scores, and diagnostic use of the Geriatric Depression Scale–Short Form (GDS-SF) over time with a sample of 177 English-speaking U.S. elders. Design and Methods: An item response theory, Rasch analysis, was conducted with data from a larger study that implemented 3 interventions and collected data at 3 time points (prior to intervention, after intervention, and a year following the end of intervention). Results: A unidimensional structure was found in this study to adequately fit the data. There appears to be significant differential item functioning, with up to one third of the items indicating that the items have different meanings for different groups of participants. As a screening tool, the GDS-SF is more likely to discriminate between older adults who have a “moderate” level of depression than between those with mild depression and no diagnosed depression. Implications: We would recommend that the GDS-SF not be used as the sole method of screening for depression but incorporated into other clinical knowledge related to the geriatric patient in question, given that not all items are equally functional. The GDS-SF may be less effective as a screening tool but could be better used to detect a change in moderate levels of depression. Clinicians may also benefit by identifying person responses that misfit. If an individual provides responses that are atypical, the person might be referred to provide a more in-depth assessment of mental health status.

Key Words: Depression in old age, Item response theory, Depression, Psychometrics, Geriatric psychiatry

The “baby boomer” generation will add 75 million people to the senior population in the United States during the next decade (Libow, 2005). Although health care spending in the United States rose 7.9% in 2004 to $1.9 trillion, accounting for 16% of the gross domestic product, costing roughly $6,280 per person per year (Smith, Cowan, Heffler, Catlin, & National Health Accounts Team, 2006), it is likely that the baby boomers will contribute to higher health care costs. Typically, older adults are diagnosed and treated for depression by their primary care physicians (Scogin & Shah, 2006). One potential way of reducing rising health care costs is improved depression screening and treatment by utilizing a widely used screening tool such as the 15-item Geriatric Depression Scale–Short Form (GDS-SF), which is in the public domain and cost free.

Based on clinical observations, the manifestation of depression in older populations is thought to be different than in younger populations. Yesavage and colleagues (1983) operationalized this difference, creating the Geriatric Depression Scale (GDS) to screen for depression in the older adult population. Sheikh and Yesavage (1986) hoped to address factors such as fatigue and poor concentration, which could make it difficult for older adults to stay focused while completing depression scales with the GDS-SF. To reduce the time it takes to fill out a scale, the GDS-SF retained 15 items with the highest item-total correlations from the original 100-item pool of the GDS. Because the GDS-SF items relate to subjective aspects of depression, its main function is for depression screening rather than diagnostic classification.

A review of the extant literature examining the psychometric properties of the English version of the GDS-SF: Short Form (GDS-SF)
the GDS-SF revealed that it has been validated in a variety of settings with a variety of geriatric samples such as adults older than 60 years living in nursing homes (Gerety et al., 1994), elders older than 65 years who visit their primary care physicians (D’ath, Katona, Mullan, Evans, & Katona, 1994), and geriatric inpatients with accompanying physical diseases (Shah, Phongsathorn, Bielawska, & Katona, 1996). Almeida and Almeida (1999) found that the GDS-SF is an adequate screening instrument for detecting a major depressive episode as defined by criteria in the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV; American Psychiatric Association, 1994) with outpatients older than 60 years. De Craen, Heeren, and Gussekloo (2003) found that the GDS-SF was applicable with people older than 85 years but that caution should be taken when there is a possibility of cognitive impairment. Internal consistency reliability estimates of the GDS-SF ranged from .76 to .83 (Iglesias, 2004; Van Marwijk et al., 1995). Kieffer and Reese (2002) explored reliability generalization by examining 338 studies, which utilized the GDS, GDS-SF, GDS-10, and GDS-4, finding .85 average reliability across these studies.

Although the GDS-SF has been used extensively, relatively few studies have addressed its structure and psychometric quality. Studies regarding the structure of the GDS-SF have come to conflicting conclusions when enlisting participation of different groups of elders. Friedman, Heisel, and Delavan (2005) found that a two-factor model fit best for cognitively intact but functionally impaired older adults. Brown, Woods, and Storandt (2007) also found that a two-factor model fit their data best for two samples of elders with and without dementia but that a one-factor model fit best for a sample hospitalized with depression, some with bipolar disorder. They found an invariant structure over time. In contrast, Incalzi, Cesari, Pedone, and Carbonin (2003) and Mitchell, Mathews, and Yesavage (1993) found a three-factor model to provide the best fit with samples of older medical inpatients and a community sample, respectively. Tang, Wong, Chiu, Lum, and Ungvari (2005) examined the psychometric quality of the GDS-SF by using a Rasch analysis with a community sample of 300 Hong Kong Chinese patients with pneumoconiosis. They found that the GDS-SF was unidimensional, but there was evidence of item redundancy suggesting that a shortened version would be as useful as the original version. In addition, they found that none of the items had significant differential item functioning (DIF) for groups categorized by age, education, and cognitive impairment. Limitations of Tang and colleagues’ study were that their findings may not be applicable to non-Chinese populations and patients with other medical diseases.

To our knowledge, no studies to date have assessed the psychometric properties of the English version of the GDS-SF such as construct dimensionality and invariance over time with an English-speaking sample. This type of analysis is useful in determining the GDS-SF’s applicability as a consistent invariant tool for measuring geriatric depression. According to Bond and Fox (2007), social science measures should be held to the same standards as measures in the physical sciences and hence demonstrated to be invariant. For example, in the physical sciences, if a yardstick is used to measure length, that ruler should measure length in the same manner even if different individuals are being measured. Similarly, a Rasch analysis of the GDS-SF seeks to determine the extent to which the scale acts as a uniform unbiased “yardstick” when measuring geriatric depression among geriatric populations. The purpose of this study was to examine scale dimensionality, reliability, invariance, targeting, continuity, cutoff scores, and diagnostic use of the GDS-SF over time. A Rasch analysis can contribute to an appraisal of the utility of the GDS-SF and of the stability of the scale over time. Although similar analyses can be conducted from a classical test theory (raw score) perspective, analysis via Rasch software automates some analyses and provides more detailed information in visual as well as tabular form.

**Methods**

As part of a larger study (Cox, Green, Hobart, Jang, & Seo, 2007) to assess the effects of three interventions directed to increasing the self-efficacy of care receivers, participants were assigned to one of three groups: (a) an individual intervention, (b) a group intervention, and (c) a case management comparison group. Responses to the GDS-SF were collected from these three groups of participants at three time points: Time 1, pretest; Time 2, 2 months following the last treatment session; and Time 3, 12 months following the last treatment session. The efficacy interventions did not specifically address depression or coping with depression, so significant direct effects of the intervention on depression were not anticipated. However, the GDS-SF was given to...
explore a potential link between self-efficacy and depression as well as provide evidence supporting or refuting convergent validity of the Care Receiver Self-Efficacy Scale (CRES; Cox, Green, Seo, Inaba, & Quillen, 2006). A subscale of the CRES, perceptions of dependence, was significantly positively correlated with depression in the sample. Also, when a grouping factor is present, it is typical to assess structure by group, if only to confirm invariance due to lack of group differences.

**Participants**

A total of 177 participants from the Denver, Colorado, metro area in the United States were recruited. The mean age of participants (see Table 1) in Time 1 data was 78.42 years ($SD = 9.78$), with ages ranging from 51 to 96 years. Most of the participants were women (76.8%), Caucasian (78.5%), and widowed (51.4%). The modal residential location was an independent retirement community (40.7%), with 27.1% in assisted living, 24.9% in their own home, 5.6% in a relative’s home, and 1.7% in a skilled nursing facility. Most participants self-reported their health status as fair or good (39.8% in each category), with 13.0% in poor and 7.4% in excellent health. The modal annual income category was $4,860 to $11,064 (51.7%), with income categories ranging from $0 to $4,860 (8.2%) to more than $45,000 (2.7%).

**Procedure**

Participants were selected if they required 6 hr of personal care per week with activities of daily living and were cognitively able to participate in...
interventions in an effort to obtain a sample representative of elders who struggled for ways to provide self-care and maintain control over the care process. Participants included individuals with disability related to conditions such as stroke, heart disease, osteoporosis, mild dementia, cancer, and severe arthritis. Intervention groups were equivalent within sampling error on demographic variables but differed by residence type at the $p < .05$ level. Participants were assigned to the group intervention if they could attend a group meeting outside their residence; all other research participants were randomly assigned to either the individual intervention or the case management (for further details, see Cox et al., 2006, 2007).

Analyses

Analyses were performed using Rasch model software, Winsteps version 3.62.1 (Linacre, 1999–2007). The Rasch model is a mathematical formula that specifies the form of the relationship between the persons and the items that operationalize a single trait. Specifically, the likelihood of higher scores increases as people have more of the trait and decreases as they have less of the trait, in this case geriatric depression, whereby items become more difficult to endorse. The Rasch model assumes that item responses are governed by a person’s position on the underlying trait and item difficulty. Item responses are modeled rather than sum total responses. The model makes no allowance for deliberate or unconscious deception, guessing, or any other variable that might impinge on the responses provided.

The Rasch model provides two parameter estimates: person location and item location, also referred to as person logit and item logit, respectively, where a logit is a nonlinear translation of the raw score. Equal-interval measures can be constructed using the Rasch model, where persons and items exist on a common scale. In other words, raw scores are nonlinearly transformed into position estimates for items and persons so that the data best fit the model.

To illustrate some differences between analyzing these data using the Rasch model and the standard method of comparing raw scores, the distribution statistics of GDS-SF scores were examined by time (Table 2). When using a classical test theory approach, there appears to be little difference between time points, and the distributions are somewhat skewed. When using Rasch analysis and examining logit position, the differences between time points are more apparent and the creation of interval scale scores removes much of the skewness.

Responses to GDS-SF items were dichotomous (agree/disagree). Rasch model software allows for polytomous responses as well; further, items can have different numbers of response options. When items have more than two possible responses, the response scale structure can be evaluated as well as items and persons, and estimates are provided that reflect the difficulty of each response for each item.

Analyses reported here were directed to assessing GDS-SF dimensionality, reliability, invariance, targeting, scale continuity, and use of cutoff scores for screening. Because Rasch model analysis is somewhat unfamiliar to many researchers and practitioners, brief descriptions of Rasch model concepts are included in the results, along with reference to analogous classical test theory indexes where available.

Results

Dimensionality

An assumption of the Rasch model is that the items assess a single or unidimensional construct. In this analysis, the items of the GDS-SF assessed the construct of geriatric depression. Dimensionality was examined in this study by (a) the dimensionality coefficient created by Wright (1994),
(b) overall fit of data to a one-dimensional model, 
(c) a Rasch principal components analysis of residuals, and (d) individual item fit.

The dimensionality coefficient resulted in coefficients of .94 at all three time points. According to Wright, dimensionality coefficient values above .9 indicate a unidimensional variable, whereas values below .5 may violate unidimensionality. This indicates that the GDS-SF is measuring one construct, geriatric depression, contrasting with earlier work finding multiple factors to underlie GDS-SF scores.

Overall fit was assessed by examining the global fit (outfit and infit statistics) of data to a unidimensional structure. Fit statistics are transformations of chi-square statistics. Fit statistics include the average fit of persons and items. The expected values of the mean square (MNSQ) fit indexes are 1.0 and 0.0, respectively, if the data fit the model exactly. Fit is expressed as “infit” (weighted by the distance between the person location and item location) and “outfit” (an unweighted measure). Infit is less sensitive than outfit to extreme responses. Person fit to the Rasch model is an index of whether individuals are responding to items in a consistent manner or responses are idiosyncratic or erratic. Similarly, item fit is an index of whether items function logically and provide a continuum useful for all respondents. An item may “misfit” because it is too complex, confusing, or actually measures a different construct (e.g., morale rather than depression). Fit can be expressed as both an MNSQ and a standardized value. Mean square fit values are reported here. Values for differentiating fit and misfit, although arbitrary, are not capricious and should be sufficiently flexible to allow for researcher judgment, though values between 0.5 and 1.5 are considered to be productive of measurement (Linacre, 2007).

For this sample, the average outfit MNSQs were 0.97, 1.02, and 0.97, at the three time points, respectively, indicating that there was little unexpected behavior by the sample in endorsing items that were far from the trait level. That is, people in the sample who were less depressed did not endorse items indicative of higher levels of depression and vice versa. Infit MNSQ is sensitive to unexpected answers of items close to the person’s trait level. Across the three time points, the infit statistics were 1.00, 0.98, and 1.00, respectively, suggesting that there was a unidimensional model—expected frequency of unexpected answers by the sample to questions that were close to their level of depression. The examination of the outfit and infit MNSQ statistics indicate that the data fit a unidimensional Rasch model well.

A Rasch principal components analysis of residuals was also conducted. The purpose of this analysis was not to construct variables, as it would be with exploratory factor analysis, but to determine whether a second dimension is indicated once the first dimension has been removed; hence, an analysis of residuals. For the present data at Time 1, the Rasch first dimension explained 72.3% of the variance in the data. The largest secondary dimension, “the first contrast in the residuals” explained 3.4% of the variance, and the eigenvalue of the unexplained variance in the first contrast was 1.8, less than the strength of two items, the smallest amount that could be considered a dimension. Based on Linacre’s (2007) recommendations for interpreting this type of analysis of residuals, the instrument may be considered unidimensional because the variance explained by the first dimension was larger than 60%, the eigenvalue for the first contrast was less than 3.0, and the variance explained by the first contrast in the residuals was less than 5%. Results at Times 2 and 3 were comparable, with 72.2% and 76.8%, respectively, of variance explained by the first dimension, eigenvalues of 1.9 and 1.9, and 3.4% and 3.0% of contrast-explained variance.

In the final examination of dimensionality, individual item fit was examined to see if any items misfit the Rasch model. Mean square outfit and infit statistics that are similar in interpretation to global fit statistics described previously were examined for each item. According to Linacre (2007), items that contribute little value to the measure have item MNSQ fit values above 1.5. Using this cutoff at Time 1, Item 6, “Are you afraid that something bad is going to happen to you?”, had an MNSQ outfit value of 1.67, which is outside the range of productive measure. At Time 2, Item 2, “Have you dropped many of your activities and interests?”, had an MNSQ outfit value of 2.43. Also at this time point, Item 13, “Do you feel full of energy?”, and Item 10, “Do you feel you have more problems with memory than most?”, had MNSQ outfit values of 2.05 and 1.65, respectively. At Time 3, none of the items reached the cutoff. These results suggest that some items on the GDS-SF contributed little to the construct of geriatric depression at Times 1 and 2.

The analogous classical test theory tool to examine dimensionality is exploratory or confirmatory
factor analysis, which assumes interval scale response data. In an exploratory factor analysis with the data in this study, four factors were found with eigenvalues over 1.0, although the eigenvalue for the first factor was four times that of the second factor. Once the dimensional structure has been identified, an item analysis using item-total correlations would be conducted to identify item misfit with the dimensions.

**Reliability**

Person and item separation and reliability of separation assess instrument spread across the trait continuum. Reliability of person separation is conceptually similar to Cronbach’s alpha, although the formulation differs. Reliability of person separation is computed without extreme scores, which can make its value lower than that for Cronbach’s alpha. Extreme scorers cannot be accurately located on the trait; we know that an extreme scorer has a higher (or lower) score than a nonextreme scorer but we are lacking knowledge about how extreme the individual’s score actually is. For example, an individual who responds with all negative answers to GDS-SF items reports very low levels of depression, but we cannot say how low without using more items designed for nondepressed people.

“Separation” measures the spread of both items and persons in standard error units. For an instrument to be minimally useful, separation should exceed 1.0, with higher values of separation representing greater spread of items and persons along a continuum. Lower values of separation indicate redundancy in the items and less variability of persons on the trait. To operationalize a variable with items, each item should mark a different amount of the trait, as, for instance, the way marks on a ruler form a measure of length. Separation, in turn, determines reliability. Higher separation in concert with variance in person or item position yields higher reliability. Person separation exceeded 1.0 at all three time points, with 1.70 at Time 1, 1.76 at Time 2, and 1.78 at Time 3, suggesting that the GDS-SF was a potentially useful instrument in measuring depression with this sample.

Reliabilities of person separation were .74, .76, and .76 at Times 1, 2, and 3, respectively. Cronbach’s alphas of .83, .85, and .83 for each time point were found, indicating that the GDS-SF had adequate internal consistency reliability with this sample at each administration of the measure. As noted earlier, with extreme scores excluded, values for reliability of person separation were lower than those for Cronbach’s alpha.

**Invariance**

Invariance is crucial to the usefulness of the GDS-SF in screening elders for depression consistently. Invariance means that parameters reflecting level of depression can be estimated for persons regardless of exactly which items were used and that parameter estimates for items are not merely a function of the particular sample to which the measure was given. We have invariant measures of height, weight, and temperature, for example, and lack measures of mainly health and psychologically related constructs that have been demonstrated to be invariant. Failure of invariance means we cannot compare samples because the construct means something different to each sample. Invariance of the GDS was assessed in two ways. First, GDS-SF item logit positions were correlated over time. Second, DIF was assessed by intervention group and for each time point.

Item logit position appeared to be fairly consistent across time points, with Item 11, “Do you think it is wonderful to be alive now?”, being the most difficult item to endorse across time points; and Item 13, “Do you feel full of energy?”, the easiest to endorse (see Table 3 for a complete listing of GDS-SF items). The correlations of item logit position were as follows: \( r = .98 \) (Time 1 to

<table>
<thead>
<tr>
<th>Table 3. Geriatric Depression Scale–Short Form Items by Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Are you basically satisfied with your life?</td>
</tr>
<tr>
<td>2. Have you dropped many of your activities and interests?</td>
</tr>
<tr>
<td>3. Do you feel that your life is empty?</td>
</tr>
<tr>
<td>4. Do you often get bored?</td>
</tr>
<tr>
<td>5. Are you in good spirits most of the time?</td>
</tr>
<tr>
<td>6. Are you afraid that something bad is going to happen to you?</td>
</tr>
<tr>
<td>7. Do you feel happy most of the time?</td>
</tr>
<tr>
<td>8. Do you often feel helpless?</td>
</tr>
<tr>
<td>9. Do you prefer to stay at home, rather than going out and doing new things?</td>
</tr>
<tr>
<td>10. Do you feel you have more problems with memory than most?</td>
</tr>
<tr>
<td>11. Do you think it is wonderful to be alive now?</td>
</tr>
<tr>
<td>12. Do you feel pretty worthless the way you are now?</td>
</tr>
<tr>
<td>13. Do you feel full of energy?</td>
</tr>
<tr>
<td>14. Do you feel that your situation is hopeless?</td>
</tr>
<tr>
<td>15. Do you think that most people are better off than you are?</td>
</tr>
</tbody>
</table>

Downloaded from https://academic.oup.com/gerontologist/article-abstract/49/2/262/551922 by guest on 29 December 2018
Although it was not anticipated that the construct intervention and case management participants. participants who were more mobile than individu-
their residence. This resulted in group intervention tion if they could attend a group meeting outside participants were assigned to the group interven-
treatment conditions constituted groups of populations of interest. In this examination, differ-
treatment conditions were conceptualized as sub-
tervention in health care settings, whereas the two adults who would typically receive this type of in-
thought to be representative of a group of older er. The case management comparison group was
intervention groups were compared with each oth-
was compared with two focal groups: the individu-
group to agree with, whereas others were easier. Similarly, for the individual and group treatment conditions, no clear pattern emerged.

There was an indication of some different mean-
ings of items at Time 1 and Time 2. At Time 1, pretest, there was significant (p < .05) DIF for four items (Item 1, “Are you basically satisfied with your life?”; Item 6, “Are you afraid that something bad is going to happen to you?”; Item 13, “Do you feel full of energy?”; and Item 15, “Do you think that most people are better off than you are?”). At Time 2, 2 months post-test, there was significant DIF with five items (Item 1, “Are you basically satisfied with your life?”; Item 2, “Have you dropped many activities and interests?”; Item 4, “Do you often get bored?”; Item 6, “Are you afraid that something bad is going to happen to you?”; and Item 10, “Do you feel you have more problems with memory than most?”; see Table 3 for GDS-SF items). At Time 3, 12-month follow-up, there was no significant DIF. In total, 13 of the 135 differences (3 time points × 15 items × 3 pairwise comparisons) were statistically significant at the p < .05 level. By chance, one would expect approximately six to seven significant differences, so the number found exceeded that. Even if a Hochberg (1988) correction were used, no items would be said to evidence DIF.

Approximately half of the DIF at p < .05 was found with four items that misfit overall (Item 2, “Have you dropped many activities and interests?”; Item 6, “Are you afraid that something bad is going to happen to you?”; Item 10, “Do you feel you have more problems with memory than most?”; and Item 13, “Do you feel full of energy?”). Differential item functioning likely contributed to item misfit. Items with the largest differences in logit positions were Item 1, “Are you basically satisfied with your life?” at Time 1; and Item 4, “Do you often get bored?”, at Time 2. In both of these cases, items were more difficult for the group intervention participants to agree with compared with the individual intervention. (Results of the 135 DIF tests can be obtained from the first author.)

Invariance analyses can be conducted via classical test theory by examining differences in item means by group or time, but such analyses are
greatly simplified via use of Rasch model software.

**Targeting and Scale Continuity**

Placement of items and persons on a common scale permits evaluation of scale function relative to the sample or targeting. The Rasch model software graphs person location with item location. In considering the construct of geriatric depression, GDS-SF items that attempt to measure somatic depression symptoms such as a lack of feeling energetic may be easier to endorse than items aimed at measuring affective depression components like not feeling happy. Logic in the arrangement of items indicates that a researcher understood the construct, adequately operationalized it with the items written, and successfully communicated it to respondents. Additionally, gaps in the measure can be examined to determine how adequately the scale forms a continuum in measuring the variable of interest, geriatric depression.

Simultaneous positioning of items and person responses illustrates where responses place each person with respect to items. This person–item graph is useful in three ways: (a) It can be used to determine the extent to which item locations match person locations. If the locations do not line up, the items are likely inappropriate for the persons (e.g., too easy or too hard to agree with). (b) Gaps in the measure can be detected that address scale continuity, suggesting where items might be added. (c) Item order can be reviewed to assess the validity of the measure. The scale function relative to the sample was performed by examining graphs of person location with item location shown in Figure 1 for Time 1. (Figures displaying Times 2 and 3 are available from the first author, but are very similar to Figure 1.)

By reviewing Figure 1, we can determine how well the measure was targeted to this sample. The left side of the figure represents participants who were graphed by intervention group. Based on the distribution, it appears that although part of the sample was measured well by the items, a portion of the sample distribution fell below −3.0 on the scale across time points. This suggests an absence of the measured trait, geriatric depression, in a portion of the sample.

The distribution of items along the scale had items generally appearing between logits −3.0 and +1.5, which appear on the right side of Figure 1. This indicates that the items in the GDS-SF have good measurement along that range, particularly between logit positions 0.0 and +1.5. An examination of gaps in the measure suggests that the items could profitably be created to measure higher levels of depression above logit +1.5 and lower levels of depression with items between logits −1.5 and −3.5. In Figure 1, the GDS-SF items are named by item number, with the first item labeled “GDS 1.”

Targeting is more difficult to identify with classical test theory because items and persons are reported using different scales (items with means and persons with raw total scores). Item means that are very high (or very low) would suggest those items to be less useful for the sample. And low variance of scores in the sample would suggest that the items were inappropriately targeted. Rasch software placement of items and persons on a joint visually presented figure assists in more clearly identifying overlap of persons and items.

**Cutoff Scores and GDS Diagnostic Use**

There have been numerous recommendations in the literature for a cutoff score to diagnose depression. Van Marwijk and colleagues (1995) recommended using a cutoff of 2/3 to diagnose mild depression with a general practice sample. Hermann and colleagues (1996) suggested using a cutoff of 5/6 with outpatients in a geriatric psychiatry clinic. Pomeroy, Clark, and Philip (2001) used a cutoff of 4/5 in samples of patients in rehabilitation wards in the United Kingdom. Tang and colleagues (2005) used a cutoff of 7/8 when using the Chinese version of the GDS-SF. According to Sheikh and Yesavage (1986), a score between 0 and 4 on the English version of the GDS-SF is considered to represent normal complaints depending on age and education; scores from 5 to 8 suggest mild depression; a score between 9 and 11, moderate depression; and a score from 12 to 15, severe depression. Using their recommendations, a score of 4 suggests a lack of depression, and a score of 5 is indicative of “mild depression.” A cutoff between 4 and 5 could distinguish between elders who are depressed and those who are not, 8/9 would distinguish between mild and moderate depression, and 11/12 between moderate and severe depression.

Sheikh and Yesavage (1986) intended the GDS-SF for use as a screening tool for geriatric depression rather than for diagnostic classification. We would expect to see a larger number of items at the logit positions corresponding to scores of
4 (normal) and 5 (mild depression). It should be noted that the Rasch model provides a nonlinear transformation of raw scores into logit positions so that every raw score corresponds to a logit position. The logit positions are on an interval scale, and the raw scores are not; the logit positions corresponding to raw scores of 4, 5, and 6 are −1.27, −.83, and −.43, respectively. In Figure 1, a bracket was placed roughly at the logit position for a raw score cutoff of 4. Ideally, items would cluster above and below the cutoff to provide discrimination of geriatric depression at that point via multiple items that reflect that level of depression. Instead, there are only two items (Item 9, “Do you prefer to stay at home, rather than going out and doing new things?”), and Item 10, “Do you feel you have more problems with memory than most?” that are close to the cutoff logit position range from −1.27 to −.83. The majority of items of the GDS-SF had logit positions above the cutoff range, suggesting that they were more difficult to agree with and indicating higher levels of geriatric depression. Many of the items corresponded to logit positions between 0.0 and 1.0 (raw scores of 7–10), suggesting that the GDS-SF items screen more effectively for varying levels of moderate geriatric depression rather than the cutoff between normative complaints (no depression) and mild depression.

Examining Figure 1, we can see that there are two items close to −1.27 (a raw score of 4), six items close to 0.60 (a raw score of 9), and two
items above 1.00 (raw scores of 11 and higher). If cutoffs are critical for diagnosis and referral for treatment, a larger number of items at these cutoff points would allow the GDS-SF to discriminate more precisely at these cutoffs. The distribution of items represented by this sample indicates that the GDS-SF has good item discrimination when differentiating between mild and moderate depression. But only two items are located at the point that discriminates between normal complaints and mild depression, suggesting that the GDS-SF may not have adequate items to differentiate those elders who are mildly depressed from those who are not. Based on the item distribution found with this sample, the GDS-SF seems adequate as an assessment tool in differentiating elders who suffer from mild to moderate depression. The development of items with less endorsability and more endorsability would be beneficial, allowing the GDS-SF to be used more effectively as a screening tool to differentiate between those who are suffering from depression and those who are not.

Iglesias (2004) examined 118 homebound, ill elderly persons with a mean age of 78.4 years: 40% of the sample fell within normal range, 23% were mildly depressed, and 37% had scores of 8 or higher, suggesting moderate to severely depressed levels. In comparison, the sample for this study had a higher percentage in the normal range, and less mild, moderate, or severe depression. For this sample at pretest, using a raw score cutoff of 4 for mild, 8 for moderate, and 11 for severe depression, 50.3% of the sample were not depressed, 23.2% were mildly depressed, 9.6% were moderately depressed, and 2.8% were in the severe range of depression. Using logit position scores, which transformed raw scores to an interval scale, 57.6% were not depressed, 23.7% were mildly depressed, 14.7% were moderately depressed, and 4.0% were severely depressed.

The chance-adjusted agreement, kappa, between individuals classified via the logit position and raw score cutoffs was $\kappa = .92$, with most discrepancies in categorization between raw score and logit categorizations found for the moderate-depression category.

Discussion

A unidimensional structure was found to fit the data in this study, although there were misfitting items at Times 1 and 2, which is generally consistent with Tang and colleagues’ (2005) Rasch analysis with a sample of Chinese patients. These results conflict with earlier work employing factor analysis to examine the structure of the GDS-SF finding two or three factors. An exception to finding a multidimensional structure for the GDS-SF was Brown and colleagues’ (2007) study, which found a unidimensional model the best fit for a sample of individuals who had been hospitalized for depression. Recommendations for use of the GDS-SF as a unidimensional or multidimensional measure involve differentiating underlying constructs assessed by items. Factor analysis is more likely to result in identification of multiple factors; Rasch analysis is more likely to find unidimensionality. As Brown and colleagues note, the valence of items as positively or negatively phrased affects the factor structure and can prove problematic for individuals facing decrements in cognitive complexity. Given the lack of consistency in results of factor analysis and the analyses, further work with varied samples seems indicated to clarify the structure of the GDS-SF. Also in need of resolution is whether the intent of this screening tool is to assess depression jointly with life satisfaction or to clearly assess depression. For example, Item 1, “Are you basically satisfied with your life?”, could assess life satisfaction in conjunction with feelings of worthlessness.

Some items included in the GDS-SF may have originally had high correlations with depressive symptoms, but four items (Item 2, “Have you dropped many activities and interests?”; Item 6, “Are you afraid that something bad is going to happen to you?”; Item 10, “Do you feel you have more problems with memory than most?”; and Item 13, “Do you feel full of energy?”) did not contribute to the construct of geriatric depression at some time points. This finding of noncontributing items is generally consistent with Tang and colleagues (2005); however, the individual items differed somewhat. Although both studies agreed that Item 10, “Do you feel you have more problems with memory than most?”, did not contribute to measuring a single underlying construct, our study found that Items 2, “Have you dropped many of your activities and interests?”; 6, “Are you afraid that something bad is going to happen to you?”; and 13, “Do you feel full of energy?”, did not contribute. Tang and colleagues found that Items 3, “Do you feel that your life is empty?”, 4, “Do you often get bored?”; and 9, “Do you prefer to stay at home, rather than going out and doing new things?”, did not contribute significantly.
Brown and colleagues (2007) also found Items 10 and 13 to contribute marginally to factors found in an exploratory factor analysis. Similarly, Jang, Small, and Haley (2001), using exploratory factor analysis, extracted a four-factor solution, with Item 13, “Do you feel full of energy?”, and Item 10, “Do you feel you have more problems with memory than most?”, loading on an unnamed factor due to incongruence among items. Item 2, “Have you dropped many of your activities and interests?”, and Item 6, “Are you afraid that something bad is going to happen to you?”, did not appear to load consistently on factors. Brown and colleagues found the items to load on a factor they called “general depressive affect,” whereas Jang and colleagues (2001) found the items to load on a factor named “pessimistic outlook.” Because the original intent of the GDS-SF was to use questions that a geriatric sample would not respond to in a defensive manner or that would impede the development of rapport, it is possible that some items were worded to be considered elder friendly and that the wording made them less clear or less clearly related to depression. In fact, of the misfitting items, three were negatively worded (Item 2, “Have you dropped many activities and interests?”; Item 6, “Are you afraid that something bad is going to happen to you?”; and Item 10, “Do you feel you have more problems with memory than most?”), and the fourth (Item 13, “Do you feel full of energy?”) was an item that failed to contribute in prior factor analyses. Rewording these questions to better reflect underlying depressive symptoms might improve these items’ contribution to the construct and their consistency across different administrations. For example, Item 2, “Have you dropped many activities and interests?”, appears to assess anhedonia and could be reworded as “Do you have fun in activities and interests?” Alternatively, these items might be replaced altogether in favor of items clearly related to depression in a group of elders with serious intractable disabilities. Item 2, “Have you dropped many activities and interests?”, may make less sense when asked of a 90-year-old recently bed bound due to a severe stroke than a 65-year-old anticipating retirement. Items and the scale in total may require modification for use with the old-old population or for groups with severe illness or disability. An example could be to modify Item 2 to ask “Are you able to have fun?”

There appears to be some DIF with up to one third of the items at Time 2, indicating that the items have different meanings for different groups of participants between the three intervention groups. This finding differs from Tang and colleagues’ (2005) finding that there was no DIF for age, education, and cognitive impairment. Bond and Fox (2007) recommended that the items revealing DIF should be investigated closely to determine what can be inferred about the underlying construct. An item may display DIF when individuals from different groups, who are assumed to be equal on the measured underlying attribute, have different probabilities of making a positive response (Stark et al., 2001). At Time 1, there were six significant DIF contrasts. In this study, one would have expected no DIF at Time 1 (baseline) and increasing evidence of DIF at Times 2 and 3 if individual items were affected differently by the three treatments. The pattern found was the reverse, with DIF at Times 1 and 2 and no evidence of DIF at Time 3. It is possible that another potential explanatory variable affected initial understanding of the construct. For these data, residence type was the only demographic variable collected that was statistically significantly associated with treatment group at pretest (Cox et al., 2006, 2007). However, differences were not found in this study by residence type, but it may be worth investigating in the future. Assignment to treatment in this study resulted in more participants than expected coming from retirement communities in the group-delivered intervention; more participants than expected living at home in the individually delivered intervention; and more participants than expected living in a facility (assisted living or skilled nursing facility) in the case management condition. It is possible that individuals attending the group intervention, who were by necessity more mobile, were less likely to agree with some GDS-SF items as a function of the items’ relationships with mobility. However, no significant differences in item means were found by treatment. But just as the factor structure of the GDS-SF seems to potentially be influenced by background characteristics of the individuals in the sample such as cognitive impairment or prior hospitalization, item placement on the underlying continuum could possibly be influenced by living arrangements as well. Although no such results were found here, disability that imposes serious limitations on an individual’s mobility may create a distinct subpopulation within the elderly population, just as populations of individuals older than 85 years may be distinct from 65-year-olds. Further work relating GDS-SF structure for groups that
differ on scores on a disability screening tool and availability of norms by age group would help to resolve this issue.

The GDS-SF was originally designed as a depression screening tool to address unique issues in assessing depression among older adults. We would expect all items to address depression rather than add constructs and to see multiple items around the cutoff between “normal” and mild depression. Based on the item distribution found with this sample, the GDS-SF seems adequate as an assessment tool in differentiating elders who suffer from mild to moderate depression. This suggests that when the GDS-SF is used as a screening tool, it is more likely to identify older adults who have a moderate level of depression than those with mild or no diagnosed depression. The addition of items closer to the 4/5 raw score cutoff could facilitate better discrimination of geriatric depression and could transform the measure into a more effective screening tool. Additionally, the development of items with less endorsability and more endorsability would be beneficial, allowing the GDS-SF to be used more effectively to differentiate between those who are suffering from depression and those who are not. In addition, to potentially assess a fuller range of depression, items with different levels of endorsability could be created. Although the DSM-IV Text Revision (2000) lists a certain number of symptoms to fulfill criteria for a major depressive episode, newly created items aimed at reflecting these criteria would require creation, validation, and further studies to assess their difficulty levels. Based upon our findings from this study, Item 11, “Do you think it is wonderful to be alive now?”, was the most difficult item to endorse; and Item 13 “Do you feel full of energy?”, was the easiest to endorse.

These findings beg the construct-validity question of whether we should consider geriatric depression to be unidimensional, a single underlying construct; or multidimensional, comprising more than one construct. Clark and Watson (1995) argue that a primary goal of scale development is to create a valid measure of an underlying construct. Cronbach and Meehl (1955) define a construct as “a postulated attribute of people, assumed to be reflected in test performance” (p. 283). Loevinger (1957) expressed the dilemma of such an undertaking, stating that “the problem of test homogeneity [should] be viewed in intimate relation with what is known of the complex causation of behavior, and of the fact that traits are manifest in a multiple, alternative, and at times dialectically opposed manner” (p. 645). When it comes to measuring the construct of depression, indexes such as the Beck Depression Inventory (BDI), which Bourman and Kok (1987) found to consist of three unidimensional scales, can be used, as can scales such as the Center for Epidemiological Studies–Depression (CES-D), which Stansbury, Ried, and Velozo (2006) found to be generally unidimensional, with the exception of positive affect items that could reflect a different construct. It appears that Yesavage and colleagues’ (1983) original intent in creating the 30-item GDS was to adequately measure depression in the “aged” population, which was thought to present unique problems to researchers and clinicians. With the creation of the GDS-SF, Sheikh and Yesavage (1986) sought to shorten the scale to address issues of fatigue and poor concentration. We argue that the construct of “geriatric depression” should be treated as unidimensional. If not, then as Clark and Watson suggest, in developing a construct, various levels of hierarchy in which the construct is embedded need to be identified. If the GDS-SF is indeed multidimensional, contrary to our results are the dimensions reflective of a higher-order construct such as depression over age 60.

Limitations of this study entail a relatively small sample that was not randomly assigned to all treatment conditions. In addition, our findings may be applicable solely to English-speaking elders in the United States. An advantage of the study was enlistment of a group that was studied over a 15-month period, allowing assessment of structure over time. However, longitudinal work over a longer time period with documentation of changes in cognitive and disability status would permit questions of invariance and clarity of the construct assessed to be addressed more fully.

Implications for Clinical Use and for Future Research

Challenges for clinicians who use the GDS-SF are to understand the usefulness and limitations of the measure so that they are effective in detecting depression in elderly persons. Similar to practices used in psychological assessment, we would recommend that the GDS-SF not be used as the sole method of screening for depression but incorporated into other clinical knowledge related to the geriatric patient in question, given that not all items are equally functional. Items that appear to
measure DSM-IV (1994) criteria for a major depressive episode, such as Item 2, “Have you dropped many activities and interests?” for anhedonia, may function differently depending on the ability status of the patient. When it is used to detect depression, clinicians need to be aware that the measure is not as accurate at lower levels of depression as it is at higher levels. Thus, the GDS-SF may be less effective as a screening tool but could be better used to detect a change in moderate levels of depression. For example, it could be utilized to determine changes in level of depression in response to psychotherapy or use of antidepressant medication during a follow-up visit, or both. Clinicians may also benefit by identifying person responses that misfit. If an individual provides responses that are atypical, the person might be referred to provide a more in-depth assessment of mental health status.

In the future, an “item bank,” possibly using all 30 items from the original GDS, could be created. Items are “banked” by calculating item logit position and conducting DIF analyses via repeated testing. Items with no or minimal DIF are banked. Measure administrators can then use a smaller number of items selected by logit position to narrow in on an individual’s level of depression. Although one would never use only one or two items to assess depression, the number of items used can be reduced to possibly as few as five while retaining quality in assessment. Also, repeated individual assessment is facilitated when one can draw from a bank of items. A process similar to this is used with computerized administration of tests in health settings such as outcomes measurement in rehabilitation (Ware, Gandek, Sinclair, & Bjorner, 2005) to minimize testing time while maintaining measurement precision. Along with banked items, automated scoring could be instituted with individualized score reports produced.

References


Received December 17, 2007
Accepted May 28, 2008
Decision Editor: William J. McAuley, PhD