International Psychogeriatrics (2015), 27:7, 1099–1111 © International Psychogeriatric Association 2014
doi:10.1017/S1041610214000210

Geriatric Anxiety Scale: item response theory analysis, differential item functioning, and creation of a ten-item short form (GAS-10)

Anne E. Mueller,[1] Daniel L. Segal,[2] Brandon Gavett,[2] Meghan A. Marty,[3] Brian Yochim,[2,4] Andrea June[5] and Frederick L. Coolidge[2]

[1] *Veterans Affairs Puget Sound Health Care System, American Lake Division, Washington, USA*
[2] *Department of Psychology, University of Colorado at Colorado Springs, Colorado, USA*
[3] *Transitions Professional Center, Portland, Oregon, USA*
[4] *VA Palo Alto Health Care System, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California, USA*
[5] *Department of Psychological Science, Central Connecticut State University, New Britain, Connecticut, USA*

ABSTRACT

Background: The Geriatric Anxiety Scale (GAS; Segal *et al.* (Segal, D. L., June, A., Payne, M., Coolidge, F. L. and Yochim, B. (2010). *Journal of Anxiety Disorders*, 24, 709–714. doi:10.1016/j.janxdis.2010.05.002) is a self-report measure of anxiety that was designed to address unique issues associated with anxiety assessment in older adults. This study is the first to use item response theory (IRT) to examine the psychometric properties of a measure of anxiety in older adults.

Method: A large sample of older adults (n = 581; mean age = 72.32 years, SD = 7.64 years, range = 60 to 96 years; 64% women; 88% European American) completed the GAS. IRT properties were examined. The presence of differential item functioning (DIF) or measurement bias by age and sex was assessed, and a ten-item short form of the GAS (called the GAS-10) was created.

Results: All GAS items had discrimination parameters of 1.07 or greater. Items from the somatic subscale tended to have lower discrimination parameters than items on the cognitive or affective subscales. Two items were flagged for DIF, but the impact of the DIF was negligible. Women scored significantly higher than men on the GAS and its subscales. Participants in the young-old group (60 to 79 years old) scored significantly higher on the cognitive subscale than participants in the old-old group (80 years old and older).

Conclusions: Results from the IRT analyses indicated that the GAS and GAS-10 have strong psychometric properties among older adults. We conclude by discussing implications and future research directions.

Key words: geriatric, anxiety assessment, anxiety, item response theory, short form

Introduction

Anxiety disorders in older adults are common, with prevalence estimates ranging from 3.2% to 14.2% depending upon diagnostic criteria and age cut-off (i.e., 55 years old and older versus 65 years old and older; Wolitzky-Taylor *et al.*, 2010). Sub-syndromal anxiety symptoms in late life are even more prevalent than formal anxiety disorders, with prevalence estimates ranging from 15% to 52.3% in community samples (Bryant *et al.*, 2008). There is a unique set of challenges associated with the assessment of anxiety in late life, including the co-occurrence of anxiety with physical health problems (e.g., Brock *et al.*, 2011; Murphy *et al.*, 2012), with other mental health problems such as depression (e.g., Cairney *et al.*, 2008), and with cognitive impairment (e.g., Yochim *et al.*, 2013). Such challenges highlight the necessity to use screening and assessment tools that are specifically designed for older adults and are well validated among older adult populations. Frequently used measures of anxiety have a number of limitations, which may restrict their applicability for use with older adults and increase the risk of misdiagnosing anxiety in this population (Edelstein *et al.*, 2008; Therrien and Hunsley, 2011). Such limitations include poor psychometric properties within the older adult population, especially poor content validity for measures that were not designed specifically for older adults and limited concurrent, predictive, and

Correspondence should be addressed to: Daniel L. Segal, PhD, Department of Psychology, University of Colorado at Colorado Springs, Colorado Springs, CO 80918, USA. Phone: +719-255-4176; Fax: +719-255-4166. Email: dsegal@uccs.edu. Received 14 Nov 2013; revision requested 23 Dec 2013; revised version received 16 Jan 2014; accepted 22 Jan 2014. First published online 27 February 2014.

construct validity. Yet another common limitation is the over inclusion of somatic items on many measures of anxiety. This may lead to inflated scores on anxiety measures among older adult respondents, especially when such symptoms are actually due to health problems.

The Geriatric Anxiety Scale (GAS; Segal *et al.*, 2010) is a self-report assessment tool designed specifically for use with older adults. Indeed, during development of the GAS, aspects of anxiety with particular salience for older adults were selected for inclusion in the measure. Preliminary studies have suggested that the GAS has good psychometric properties in clinical and community samples of older adults (Segal *et al.*, 2010; Yochim *et al.*, 2011). The purpose of the current study was to investigate the psychometric properties of the GAS using item response theory (IRT) and to create a short form of the measure. Short forms of screening measures are preferable in busy clinical settings and in lengthy research protocols to reduce the burden of administration time and scoring. Short forms may also be especially beneficial for some older respondents to reduce possible fatigue from longer assessments. Under an IRT model, short forms can be equally or more reliable than full-length forms, which is why IRT was used in the present study.

Research has identified age and sex as variables that differentially impact the incidence of anxiety in late life. These variables are important to consider in anxiety assessment as they may be useful in identifying individuals who are at risk for experiencing this condition. Although anxiety disorders are common among older adults, they are generally reported as less prevalent in older adults than in younger adults (Gum *et al.*, 2009; Flint *et al.*, 2010). Women tend to report higher levels of anxiety than men (Gum *et al.*, 2009) as well as more chronic anxiety (De Beurs *et al.*, 2000). Such variations could be the result of true differences in prevalence rates, but could also reflect measurement bias.

Measurement bias occurs when a particular group of individuals has an unequal chance of endorsing an item than another group of individuals despite being matched upon the variable of interest. For example, men and women with the same level of anxiety should have the same likelihood of endorsing a particular item on a measure in the same manner. Thus, if an assessment tool is biased against sex, sex-based differences in prevalence rates would reflect this measurement error instead of actual group differences. If a measure has no detectable bias but differences between groups remain, then the differences are more likely to reflect actual variations between groups. Several researchers have found item biases in various measures of anxiety (e.g., Leach *et al.*, 2008a; Van Dam, *et al.*, 2009). Thus, although age and sex appear to have an impact upon the incidence of anxiety in late life, item bias must be taken into account in scale construction.

Item response theory is a set of statistical models used to measure latent variables (e.g., anxiety), and posits that responses on a given item are a function of both person and item properties. According to IRT, individuals who have a greater level of the latent trait should have a higher probability of endorsing a particular item measuring that trait. Analyses are often presented as item characteristic curves (ICCs), plots that indicate the likelihood of endorsing an item (e.g., symptom) as the level of the underlying trait (e.g., anxiety) changes. The underlying trait is represented as theta (θ). Steeper slopes in ICCs indicate that the item under scrutiny is better able to discriminate among people with high or low levels of the latent trait (represented as the discrimination parameter; a). The threshold parameter (also known as item severity or difficulty parameter; b) indicates the trait level at which the likelihood of endorsing a given response choice is 50%. A higher threshold parameter indicates that the individual must have higher levels of the latent trait to have a 50% likelihood of endorsing the response choice. Each item has an information function, depicted in item information curves (IICs), which provide data about how much information the item yields about the latent trait. In IICs, a steeper slope indicates that the item provides more information about the latent trait, but over a more restricted range. A less steep slope indicates that the item provides less information over a more broad range. Item information functions are combined to create a test information function (TIF).

Within the IRT framework, the reliability of a test increases with the inclusion of better or more informative items. Standard error is determined by calculating the square root of the inverse of information. Both information and standard error are believed to vary across all trait levels such that a particular item or sum of items may be more informative for an individual with a higher level of a trait compared with an individual with a lower level of the trait. Thus, IRT analyses are pertinent to scale development, as items that do not provide reliable information about a person's standing on a latent trait can be identified and either rewritten or removed from the measure. IRT can also be used to analyze item bias or measurement invariance (also known as differential item functioning (DIF); Pedraza and Mungas, 2008). If an item is biased against a certain group characteristic (e.g., age and sex), then the ICCs for that item will differ despite the groups being matched on the latent trait.

Item response theory offers several notable advantages over classical test theory (CTT; de Ayala, 2009). One main advantage is that IRT analyses are not test- or sample-dependent as in CTT, resulting in more meaningful parameter estimates. Since IRT is not test- or sample-dependent, the item parameter estimates from the current study can be generalized beyond the sample used. Another advantage is that IRT has stronger assumptions than CTT, meaning that the assumptions of unidimensionality and local independence are more difficult to meet than the assumptions in CTT. Stronger assumptions beget stronger findings; thus, the use of IRT is considered more advanced than CTT techniques.

The GAS (Segal *et al.*, 2010) is a 30-item self-report measure of anxiety symptoms designed for use with older adults. There are 25 scorable items that assess symptoms of anxiety and five additional items that assess common content areas of worry among older adults. The scale contains three conceptually based subscales tapping into various components of anxiety symptoms: somatic, cognitive, and affective. This scoring format offers advantages over other measures of anxiety, as the clinician or researcher can easily determine which types of symptoms are more problematic for the respondent. In addition, if a respondent scores very highly on the somatic subscale, this raises attention to the possibility of comorbid medical conditions inflating scores on the subscale. Segal *et al.* (2010) examined the validity and internal consistency of the GAS in both community-dwelling and clinical samples of older adults. The researchers found the GAS to have excellent internal consistency in both samples. The measure also demonstrated evidence of convergent validity in its significant correlations with other measures of anxiety. Furthermore, neither the GAS total score nor the subscales correlated significantly with education, additional evidence of discriminant validity. These researchers concluded that the GAS demonstrated strong preliminary evidence for convergent and divergent validity as well as reliability. The study did not examine the impact of age or sex on measurement bias. In addition, the researchers noted that the sample sizes were not large enough to conduct factor analyses to examine the underlying factor structure of the GAS, and thus the subscales remained conceptually designed instead of empirically based.

Yochim *et al.* (2011) further examined the psychometric properties of the GAS in a community-dwelling sample of 117 older adults. As in Segal *et al.* (2010), the GAS was found to have excellent internal consistency ($\alpha = 0.90$), and significantly correlated with other measures of anxiety (evidence of convergent validity). The GAS correlated weakly with reading ability and processing speed, suggesting discriminant validity. Furthermore, Yochim *et al.* found that the GAS total score correlated with self-reported medical burden, as did the Beck Anxiety Inventory. Not surprisingly, the somatic subscale correlated with medical burden most strongly, although the affective and cognitive subscales correlated with medical burden as well. The GAS total score was not significantly correlated with sex. The correlations between the GAS subscales and sex were not reported. Factor analysis was not conducted, nor was measurement bias assessed.

The present study was the first to use IRT to examine the scale properties of an anxiety measure in an older population (Aim 1). As the GAS is intended to be a clinically useful measure of anxiety, it was expected that items should be able to discriminate individuals with high and low levels of anxiety. Items that provided low or very low levels of information were considered for removal from the scale. It was also expected that the peak of the test information curve (TIC) would be above the mean level of anxiety. The presence of DIF by age and sex was also assessed (Aim 2). This aim was exploratory in nature and no specific hypothesis was generated. A short form was created by identifying and retaining the items that provided the greatest information and had the highest discrimination parameters while maintaining the integrity of the subscales (Aim 3). It was expected that the short form would have adequate reliability and function similar to the full version of the GAS. Age and sex differences were assessed at the group level, with the expectation that individuals who are younger and female would score higher on the GAS than individuals who are older and male (Aim 4).

Method

Participants and procedures

Four existing datasets were combined, yielding a total sample of 581 community-dwelling adults aged 60 years and older (mean age (M) = 72.32 years, SD = 7.64 years, range = 60 to 96 years). The sample predominately consisted of women ($n = 372$, 64%) and was primarily European American/White ($n = 511$, 88%). Other demographic information is presented in Table 1. Trait heterogeneity is recommended in estimating polytomous IRT models (i.e., used for scales with more than two response options; Embretson and Reise, 2000). The large sample size provided adequate trait heterogeneity for the proposed analyses. All participants provided informed consent as part

Table 1. Means, standard deviations, and ranges for all demographic information and measures

		N	MEAN	SD	POSSIBLE RANGE	RANGE
Sample 1	Age (years)	407	73.78	7.14	–	60–96
	Education (years)	405	14.98	2.95	–	8–25
	GAS total scale	384	9.18	7.88	0–75	0–48
	GAS cognitive	398	1.92	2.64	0–27	0–17
	GAS affective	395	2.40	2.75	0–24	0–17
	GAS somatic	398	5.07	3.63	0–24	0–20
Sample 2	Age (years)	136	68.60	7.60	–	60–88
	Education (years)	119	14.61	2.84	–	6–28
	GAS total scale	121	19.95	11.50	0–75	0–54
	GAS cognitive	128	5.91	4.03	0–24	0–17
	GAS affective	127	6.10	3.66	0–24	0–16
	GAS somatic	131	8.01	4.70	0–27	0–26
Sample 3	Age (years)	38	69.92	7.98	–	60–90
	Education (years)	38	13.95	2.10	–	9–18
	GAS total scale	38	15.16	11.12	0–75	1–47
	GAS cognitive	38	3.70	4.12	0–27	0–18
	GAS affective	38	4.32	3.88	0–24	0–14
	GAS somatic	38	7.05	4.17	0–24	1–16
Sample 4	Age (years)	581	72.32	7.64	–	60–96
	Education (years)	562	14.83	2.89	–	6–28
	GAS total scale	541	11.99	10.11	0–75	0–54
	GAS cognitive	562	2.94	3.53	0–24	0–18
	GAS affective	559	3.37	3.43	0–24	0–17
	GAS somatic	565	5.87	4.13	0–27	0–26
	GAS-10	563	4.72	4.64	0–30	0–24

Note. GAS = Geriatric Anxiety Scale, GAS-10 = Geriatric Anxiety Scale – 10-item version.

of an institutional review board (IRB)-approved research protocol prior to participation.

Sample 1. Data were collected from 123 older adults. Participants were volunteers from the community who participated in a larger study of cognitive functioning and mental health. They were tested in a private, quiet room in a university research setting. All participants were financially compensated for their time.

Sample 2. Data were collected on 284 community-dwelling older adults recruited from the El Paso county voter registry. Participants completed a mail-in packet of questionnaires.

Sample 3. Data were collected on 136 older adults receiving psychological services at a local community outpatient mental health clinic. Participants in this sample were administered measures at intake.

Sample 4. Data were collected from 38 community-dwelling older adults with at least one chronic physical health condition. Participants were tested in a quiet, private research room, and were administered a packet of mental health questionnaires. All participants were compensated financially for their time.

Measures

Geriatric Anxiety Scale

The GAS (Segal *et al.*, 2010) contains 25 self-report items used for scoring as well as five additional items that tap into common topical concerns of anxiety among older adults, for example, worry about becoming a burden to one's children. Participants are asked to rate symptoms of anxiety or stress by indicating how often they have experienced each symptom during the past week on a Likert-type scale that ranges from 0 (*not at all*) to 3 (*all of the time*). Possible scores range from 0 to 75, with higher scores indicating the presence of more severe anxiety.

Statistical analyses

The data were prepared for IRT by collapsing response categories that were infrequently endorsed ("sparse cells"). The graded response model (GRM; Samejima, 1969) was used for IRT analyses (Aim 1). There are two assumptions for IRT: unidimensionality (i.e., there is only one underlying factor within the data) and local independence (items should not be correlated when the shared variance of the latent trait is removed).

Both assumptions were tested by conducting a confirmatory factor analysis using Mplus.

All IRT analyses were conducted in R version 2.13.2 (R Core Team, 2012). Information curves, ICCs, threshold parameters, and discrimination parameters were analyzed to examine the item properties of the measure and identify which items were more or less useful in reliably measuring trait levels of anxiety. The TIC was examined to determine at what level of anxiety the GAS provides the most information. Baker (2001) suggests that discrimination parameter values ranging from 0.01 to 0.24 are considered very low, 0.25 to 0.64 are low, 0.65 to 1.34 are moderate, 1.35 to 1.69 are high, and more than 1.7 are very high. These criteria were used to determine which items were best able to discriminate among levels of anxiety.

Within the GRM, there are $k - 1$ threshold parameters (k indicates the number of response categories). As the GAS has four possible response categories (not at all, sometimes, most of the time, all the time), there are three threshold parameters presented for each item in this study. The first threshold parameter (scaled as a z-score, M = 0, SD = 1, lower values reflecting less anxiety) reflects how much anxiety is required to have a 50% chance of endorsing the "sometimes" response category. The second threshold parameter reflects how much anxiety is needed to have a 50% chance of endorsing the "most of the time" category, and the third threshold parameter reflects how much anxiety is needed to have a 50% chance of endorsing the "all of the time" category. Response options with very low or negative threshold parameters would be considered less useful items in measuring anxiety, as very low levels of anxiety would be needed to endorse these response options. In contrast, response options with extremely high threshold parameters would also be less useful, as extreme or atypical levels of anxiety would be needed to endorse these response options.

Analyses were also conducted to identify items that may demonstrate DIF with regard to age and sex (Aim 2). Age was dichotomized into young-old (60–79 years old; $n = 461$) and old-old (80 years old and older; $n = 119$). Sex was dichotomized as male or female. Logistic regression was used to investigate DIF based on the latent anxiety estimates derived from IRT using the "lordif" package in R (Crane *et al.*, 2006; Choi *et al.*, 2011).

A short form was created by identifying ten items with the highest discrimination parameters while retaining the structure of the subscales, and the reliability of the short form was tested (Aim 3). It was expected that the short form would have similar properties as the full version of the scale. The items were then examined to ensure there were not redundant or similar items included in the short form, and the threshold parameters were inspected to determine if the parameters were reasonable in magnitude (i.e., not extremely low or high). The procedure for the selection of the short form items followed Edelen and Reeve (2007). The TIC for the short form was then inspected to ensure that the short form provided a reasonable level of information in comparison with the full version.

Both t-tests and analysis of variance (ANOVA) were conducted on the data to test the hypotheses that younger, female individuals would report the highest levels of anxiety (Aim 4).

Results

Aim 1: Confirmatory factor analysis and IRT. Confirmatory factor analysis suggested a unidimensional model provided acceptable fit to the data, $\chi^2(275) = 1582.97$, p = 0.00, CFI = 0.923. TLI = 0.916, and RMSEA = 0.091 (90% CI [0.086, 0.095]). It should be noted that the χ^2 test is inflated due to the large sample size. The CFI, TLI, and RMSEA values were close to acceptable. Three pairs of items had high modification indices, indicating there may be additional covariance shared by these pairs of items after controlling for anxiety. These item pairs included the following: Items 6 ("I was afraid of being judged by others") and 7 ("I was afraid of being humiliated or embarrassed"), 8 ("I had difficulty falling asleep") and 9 ("I had difficulty staying asleep"), and 18 ("I worried too much") and 19 ("I could not control my worry"). Removal of any of these items would increase the fit of the unidimensional model to the data. Based upon the redundancy of items 8 and 9, it was decided that item 9 would be removed from the analysis. Also, although the other item pairs are similar in content, they measure separate symptoms of anxiety (i.e., excessive worry and difficulty controlling worry are two separate symptom criteria of Generalized Anxiety Disorder). Furthermore, the magnitude of the modification index between items 8 and 9 was the strongest, suggesting the largest potential for improved fit. CFA was performed on the remaining 24 GAS items and revealed a better fit to the data, $\chi^2(252) = 1051.17$, p = 0.00, RMSEA = 0.074 (90% C.I. [0.069, 0.079]), CFI = 0.951, and TLI = 0.947. The results from this subsequent CFA were deemed sufficient to meet the statistical assumption of unidimensionality for the IRT analyses. Item 9 was excluded from the IRT analyses.

The discrimination and threshold parameters for all GAS items are listed in Table 2. IRT analyses yielded discrimination parameters ranging from

Table 2. IRT calibration for GAS items

GAS ITEM	DISCRIMINATION a	THRESHOLD b_1	THRESHOLD b_2	THRESHOLD b_3
Item 1 (My heart raced or beat strongly)	1.160	0.689	3.558	4.526
Item 2 (My breath was short)	1.070	0.542	3.258	5.094
Item 3 (I had an upset stomach)	1.086	0.878	3.508	4.908
Item 4 (I felt like things were not real or like I was outside of myself)	1.937	1.670	3.248	–
Item 5 (I felt like I was losing control)	2.473	0.983	2.440	–
Item 6 (I was afraid of being judged by others)	1.516	0.767	2.870	4.013
Item 7 (I was afraid of being humiliated or embarrassed)	1.396	1.031	3.251	–
Item 8 (I had difficulty falling asleep)	1.067	–0.062	2.108	3.392
Item 10 (I was irritable)	2.026	–0.057	2.458	3.368
Item 11 (I had outbursts of anger)	1.466	0.762	3.582	–
Item 12 (I had difficulty in concentrating)	1.987	0.090	2.185	3.297
Item 13 (I was easily startled or upset)	2.041	0.643	2.425	–
Item 14 (I was less interested in doing something I typically enjoy)	2.172	0.438	2.151	2.985
Item 15 (I felt detached or isolated from others)	2.226	0.765	1.978	2.896
Item 16 (I felt like I was in a daze)	2.214	1.180	2.740	–
Item 17 (I had a hard time sitting still)	1.360	0.876	2.651	4.032
Item 18 (I worried too much)	2.225	–0.072	1.568	2.535
Item 19 (I could not control my worry)	2.657	0.569	1.802	2.789
Item 20 (I felt restless, keyed up, or on edge)	2.657	0.296	1.998	2.848
Item 21 (I felt tired)	1.758	–0.993	1.410	2.456
Item 22 (My muscles were tense)	2.040	0.040	1.839	2.805
Item 23 (I had back pain, neck pain, or muscle cramps)	1.171	–0.874	1.387	2.679
Item 24 (I felt like I had no control over my life)	3.024	0.754	1.818	2.481
Item 25 (I felt like something terrible was going to happen to me)	2.371	1.487	2.748	–

Note. Items without b_3 parameter estimates had been collapsed due to sparse cells. Items included on GAS-10 are highlighted in bold.

1.070 to 3.024. These parameters reflect values in the moderate to very strong range (Baker, 2001). Items from the somatic subscale (M discrimination parameter = 1.339, SD = 0.366) tended to have lower discrimination parameters than items from the cognitive (M = 1.938, SD = 0.442) and affective subscales (M = 2.361, SD = 0.358). This indicates that the somatic items were less informative than the cognitive and affective items. Threshold parameters in the current study ranged from –0.993 to 1.670 for the first parameter, 1.387 to 3.582 for the second parameter, and 2.456 to 5.094 for the third parameter.

Next, the TIC was examined. The TIC and the standard error estimate (SEE) are presented in Figure 1. The GAS provides the greatest amount of information for individuals with average or higher levels of anxiety as indicated by the maximum TIC and minimum SEE. The TIC peak was at approximately 2.5 SD above the mean level of anxiety. The TIC also indicated that the GAS provides less information above 3 SDs above the mean level of anxiety, as well as 1 or more SD below the mean level of anxiety. This is reflected by low TIC and higher SEE. This would indicate that the GAS does not provide useful or reliable information at markedly low levels of anxiety, nor does it provide useful or reliable information at levels of extremely severe anxiety.

Aim 2: Item bias (DIF). Exploratory analyses were also conducted to detect DIF among the GAS items. Item 3 ("I had an upset stomach") was flagged for age-related DIF. Test and ICCs for this item and the GAS total scale are presented in Figure 2, and indicate that the young-old group was more likely to score higher on item 3 than the old-old group with equivalent levels of anxiety. The plot on the left of this figure indicates that there

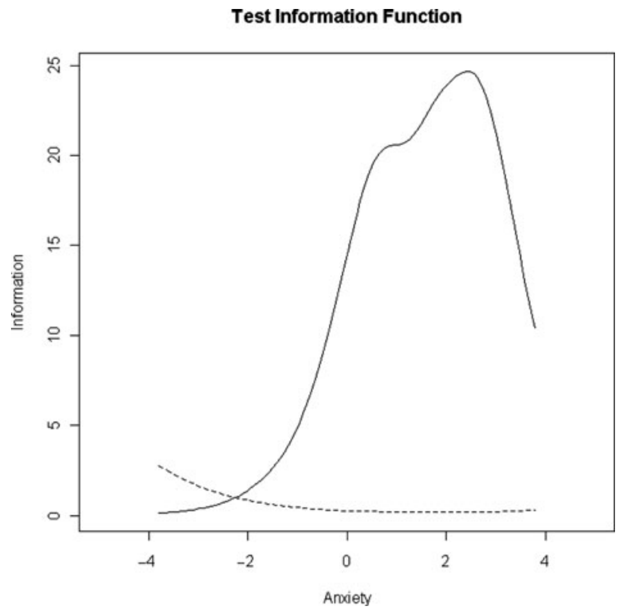

Figure 1. Test information function (TIF) for 24-item GAS. The dotted line represents standard error estimate (SEE), and the solid line represents the summative test information.

were no differences in GAS total scale scores by age, suggesting that total scale scores were not affected by DIF on item 3. An examination of McFadden's pseudo R^2 statistics indicated that the magnitude of DIF for this item was very small or negligible (0.0119; Zumbo, 1999).

Item 10 ("I was irritable") was flagged for DIF with regard to sex. The test and ICCs for this item and the total scale are presented in Figure 3, indicating that men were slightly more likely to score higher on this item than women with equivalent levels of anxiety. The plot on the left of this figure indicates that there were no differences in GAS total scale scores by sex, suggesting the amount of DIF for this item was negligible. An examination of McFadden's pseudo R^2 statistics also indicated that the magnitude of DIF for this item was negligible (0.0118; Zumbo, 1999).

Aim 3: Short form. To create the GAS short form (GAS-10), items with the highest discrimination parameters and information curve peaks were retained from each of the subscales. Three items were selected from the somatic subscale (items 17, 21, and 22) and from the affective subscale (10, 15, and 20) whereas four items were retained from the cognitive subscale (16, 19, 24, and 25). The items were examined to ensure that the short form would not contain redundant items. For instance, both items 5 ("I felt like I was losing control") and 24 ("I felt like I had no control over my life") had high discrimination parameters and information curve peaks, but both pertain to perceptions of control. In this case, the latter was retained due to the higher discrimination parameter. The threshold parameters of the ten chosen items were also inspected, and revealed items that could be endorsed by individuals with varying amounts of anxiety severity (i.e., the threshold parameters were neither too high nor too low). Presented in Figure 4, the TIC for the GAS-10 indicates the short form best assesses anxiety for individuals with average up to 2.5 SD above the mean level of anxiety. Comparable to the full version of the GAS, the GAS-10 does not assess anxiety reliably for people below the mean level or 3 SD above the mean of anxiety. The peak of the TIC indicated that a reasonable amount of information is provided by the GAS-10 in comparison to the full GAS. The SEE indicated that the GAS-10 did not lose a large amount of precision as a result of reducing items.

With regard to measures of reliability, the GAS-10 performed similar to the full version of the GAS. The GAS-10 had excellent internal consistency (Cronbach's $\alpha = 0.89$), and was significantly and positively correlated with the GAS total scale ($r = 0.96$, $p < 0.001$) and subscales (cognitive: $r = 0.92$, $p < 0.001$, affective: $r = 0.89$, $p < 0.001$, somatic: $r = 0.82$, $p < 0.001$).

Aim 4: Age and sex differences in scores. A series of independent sample t-tests were conducted to compare GAS total scale (25-item version) and subscale scores for men and women. Women ($M = 13.12$, $SD = 10.58$) scored significantly higher than men ($M = 10.02$, $SD = 8.93$) on the GAS total scale score, $t(465.73) = 3.63$, $p < 0.001$, $d = 0.31$, 95% CI [0.14, 0.48].

Women ($M = 3.37$, $SD = 3.68$) also scored higher than men ($M = 2.17$, $SD = 3.13$) on the cognitive subscale, $t(476.66) = 4.09$, $p < 0.001$, $d = 0.34$, 95% CI [0.17, 0.51]. Similarly, on the affective subscale, women ($M = 3.69$, $SD = 3.59$) scored higher than men ($M = 2.80$, $SD = 3.07$), $t(481.28) = 3.10$, $p < 0.01$, $d = 0.26$, 95% CI [0.09, 0.43]. Finally, there were sex differences on somatic subscale scores as well, such that women ($M = 6.30$, $SD = 4.26$) scored higher than men ($M = 5.10$, $SD = 3.77$), $t(563) = 3.34$, $p < 0.01$, $d = 0.29$, 95% CI [0.12, 0.46].

Next, a series of independent t-tests were conducted to examine potential age differences in GAS scores among older adults only. With regard to GAS total score, there was no statistically significant difference between the young-old ($M = 12.31$, $SD = 10.72$) and old-old ($M = 10.51$, $SD = 7.77$; $t(224.23) = -1.77$, $p = 0.08$, $d = 0.17$, 95% CI [−0.03, 0.38]). In contrast, the young-old ($M = 3.08$, $SD = 3.74$) scored significantly higher than the old-old ($M = 2.37$, $SD = 2.55$) on the cognitive subscale, $t(255.25) = -2.40$, $p < 0.05$, $d = 0.20$, 95% CI [−0.001, 0.40].

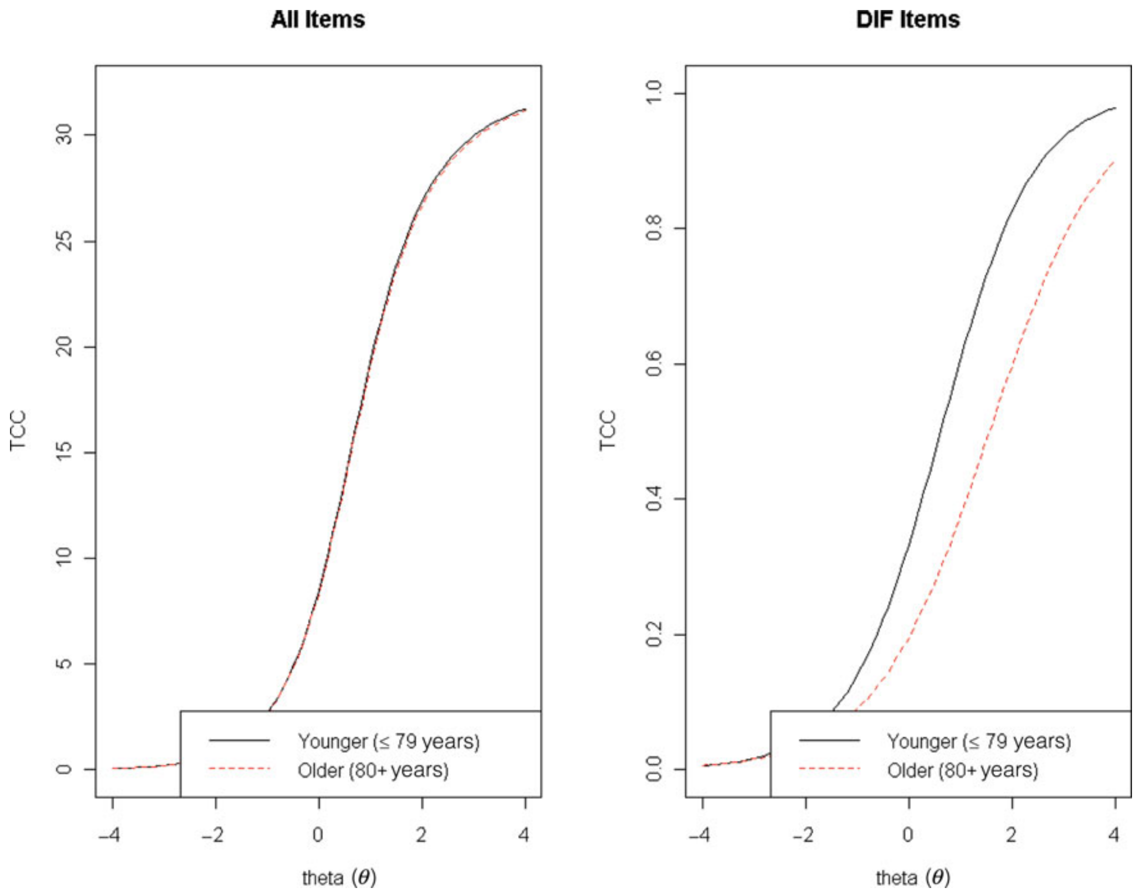

Figure 2. (Colour online) Test (left panel) and DIF-item (right panel) characteristic curves by age (young-old versus old-old).

Finally, there were no significant differences between the young-old ($M = 3.49$, $SD = 3.58$) and the old-old ($M = 2.91$, $SD = 2.75$) on the affective subscale, $t(224.95) = -1.86$, $p = 0.06$, $d = 0.17$, 95% CI [–0.03, 0.37], nor were there significant differences between the young-old ($M = 5.98$, $SD = 4.23$) and the old-old ($M = 5.46$, $SD = 3.69$) on the somatic subscale, $t(563) = -1.20$, $p = 0.23$, $d = 0.13$, 95% CI [–0.03, 0.37].

To examine the potential interaction effect between age and gender, a two-way, between-groups ANOVA was conducted. The interaction between age and sex was not statistically significant with regard to total GAS scores, $F(1, 537) = 1.46$, $p = 0.23$. Similarly, the interaction was not significant with regard to cognitive subscale scores, $F(1,558) = 2.61$, $p = 0.64$, affective subscale scores, $F(1, 555) = 1.03$, $p = 0.31$, and somatic subscale scores, $F(1, 561) = 2.62$, $p = 0.11$.

Discussion

The purpose of this study was to examine the psychometric properties of the GAS, assess for potential measurement bias, and to create a ten-item short form of the measure, called GAS-10. This study adds to the literature on anxiety assessment in several novel ways. This is the first study to utilize IRT to examine a measure of anxiety designed specifically for use with older individuals. Furthermore, this study utilized a large sample of older individuals recruited from a number of settings. Overall, the results from the current study support the use of the GAS in measuring late life anxiety. As the number of older adults who experiences anxiety increases, evidence-based assessment tools become increasingly necessary.

Item response theory analyses indicated that the GAS is most reliable in discriminating individuals at the higher end of the anxiety continuum versus people with very low levels of anxiety. As indicated by the peak of the TIC and minimum SEE, the GAS provides the most information at 2.5 SD above the mean level of anxiety. Practically speaking, clinicians are less interested in measuring anxiety in someone with below average levels of anxiety. In addition, individuals scoring 3 or more SD above the mean of anxiety would likely be experiencing extreme mental distress, which would be quite apparent in a clinical setting. Thus, the results of the IRT analysis support the use of the GAS

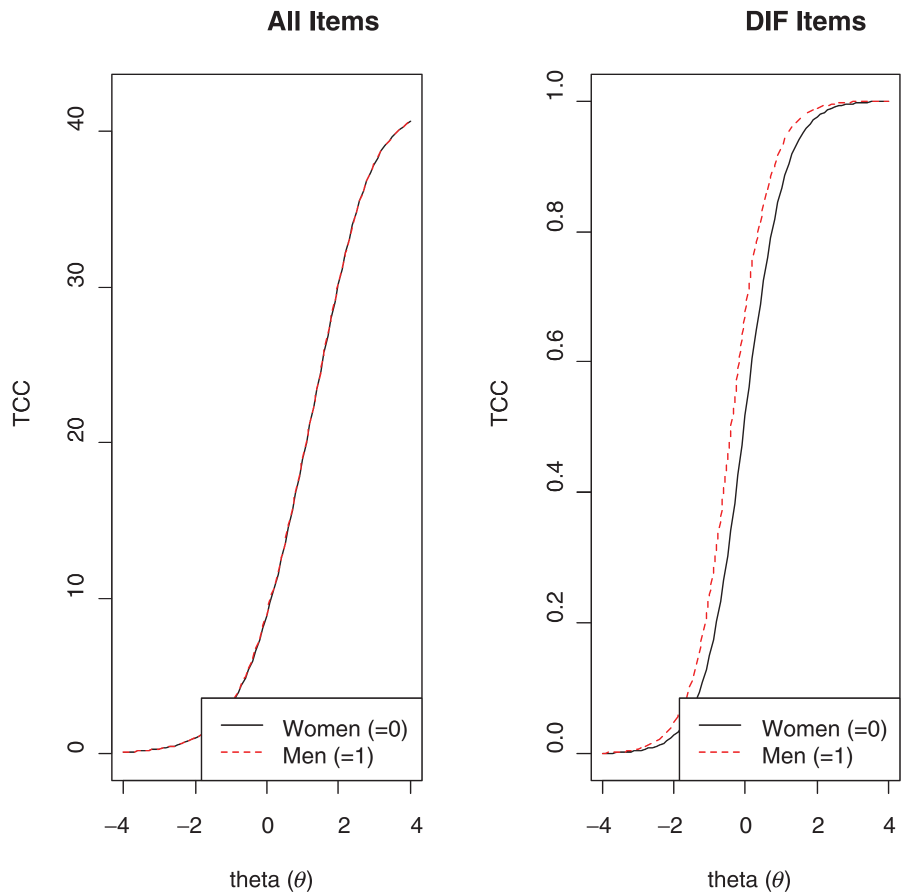

Figure 3. (Colour online) Test (left panel) and DIF-item (right panel) characteristic curves by sex.

as a clinically meaningful assessment tool for the reliable measurement of high average to above-average levels of anxiety. However, the shape of the TIC indicates that the GAS may have limitations with regard to measuring changes in anxiety scores, especially when one moves into or out of anxiety levels that are not measured well by the test (i.e., <-1 SD). This is due to the fact that at these levels of anxiety, measurement precision is low. This indicates that it would take a very large change to know that the change is reliable and not due to measurement error. Longitudinal data were not included in this study, which limits the conclusions that can be drawn regarding the longitudinal measurement properties. Clinicians and researchers should utilize the scale bearing these limitations in mind, and future studies should examine these limitations further.

The discrimination parameter values in this study ranged from moderate to very high (Baker, 2001). Overall, the somatic items provided less information than items from the affective or cognitive subscales, although all items had discrimination parameters at least in the "moderate range" ($a > 1.35$). There were no items that appeared problematic in this respect, and IRT analyses indicated that all items had justification for remaining in the scale. However, the results indicate that the somatic items provide less information about anxiety than the affective and cognitive items, likely because they are endorsed frequently by individuals with comorbid medical conditions (Katon *et al.*, 2007).

One item was flagged for DIF by age, and another item was flagged for DIF by sex. The degree of DIF for both items was very small or negligible, indicating that the overall scale scores were not impacted meaningfully by DIF. For these reasons, no modifications to these items are proposed, although researchers should continue to look closely at these two items in future studies. Women tended to score higher on the GAS total scale and subscales than men, as hypothesized, and the effect sizes for these results were small to medium (according to Cohen's (1988) conventions). This finding is consistent with previous research (De Beurs *et al.*, 2000; Lowe and Reynolds, 2005; Gum *et al.*, 2009; Potvin *et al.*, 2011), and it should be noted that the current study examined sex differences in anxiety severity versus prevalence.

Taken together with the results from the DIF analyses, the results from the t-tests indicate true group differences in anxiety severity rather than

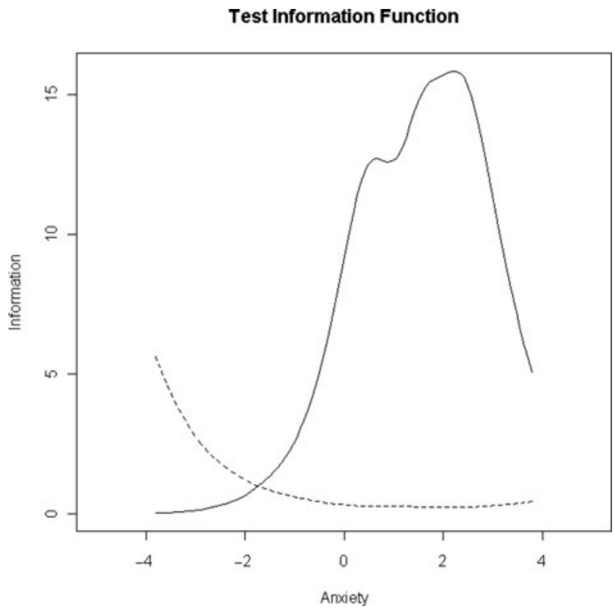

Figure 4. Test information function (TIF) for GAS-10. The dotted line represents standard error estimate (SEE), and the solid line represents the summative test information.

item bias. There are a number of reasons why women may score higher on the GAS than men. Women tend to have more risk factors for anxiety, such as poorer physical health, physical inactivity, and more interpersonal difficulties (Leach *et al.*, 2008b). Women may also be more comfortable in expressing mental health symptoms than men, leading to higher scores on self-report measures of mental health problems.

As far as age differences in the full older adult sample are concerned, the young-old (60–79 years old) scored significantly higher on the cognitive subscale than the old-old (80 years old and older). No statistically significant differences between groups were found on the total scale, affective, or somatic subscales. The hypothesis that the young-old group would report greater symptom severity than the old-old group was not supported, and contradicts previous literature reporting differences among the young-old and old-old in anxiety prevalence (e.g., Schaub and Linden, 2000; Gum *et al.*, 2009). It could be that the affective or somatic subscales are not as sensitive to age differences in anxiety symptoms among the young-old and old-old. It should be noted that there are no agreed-upon standards for classifying the young-old and old-old in the current literature, with researchers using various cut-offs. Older adults can also be split into more than two groups (i.e., young-old, middle-old, old-old, very old-old), and some use 85 years old as a cut-off for the "very old-old" (i.e., Schaub and Linden, 2000). Categorizing age in a different way could potentially have resulted in different results.

This study also created a brief version of the GAS (GAS-10), which had excellent psychometric properties with regard to reliability, convergent validity, and factor structure. The items of this short form were selected upon their discrimination parameters and information curves while retaining the structure of the subscales, following an example set forth by Edelen and Reeve (2007). While retaining only the items with the highest discrimination parameters or the highest information curve peaks may have resulted in a short form with higher information, including items from the other subscales ensured breadth of content coverage. Furthermore, the GAS-10 performed similar to the full version of the GAS with respect to the number of items retained. The detection of anxiety in primary care settings is poor (Calleo *et al.*, 2009), highlighting the dire necessity for behavioral health screening in medical settings. Short forms are also useful in epidemiological research and in clinical use with individuals with limited cognitive capacity. Thus, the GAS-10 shows promise as a brief measure of anxiety.

The CFA supported the GAS as a one-factor scale, and the model fit was improved with the removal of Item 9 ("I had difficulty staying asleep"). In this one-factor model, all items on the GAS appear to tap into the same underlying latent variable or construct (general anxiety). Future studies should explore two- and three-factor models of the GAS, with particular attention to somatic items. For instance, the Beck Anxiety Inventory contains two primary subscales (somatic and cognitive) as shown in a validation study among older adults (Kabacoff *et al.*, 1997). The Geriatric Anxiety Inventory appears to have one single underlying factor (Byrne and Pachana, 2011), which could be due to the scale's deliberate exclusion of somatic items and its content focus on worry. The GAS subscales should be utilized with these empirical findings considered. The subscales have clinical and practical utility, but clinicians should bear in mind that this study supports a one-factor model of anxiety. Future research should examine the possibility of the GAS having a bi-factor model structure. Furthermore, the possibility of the GAS lacking unidimensionality may limit the interpretation of IRT results.

As stated previously, IRT is considered more advanced than CTT techniques. One major limitation of CTT is the assumption that the test functions the same for all levels of the latent trait. The results of this study indicate that the reliability and standard error of the GAS is different for individuals among varying levels of anxiety. IRT methods can also be used to improve scoring, such as using computerized adaptive testing (CAT). One advantage of using this scoring method is that each individual has his or her own unique SEE (and thus unique confidence intervals around their estimated level of anxiety), which take into account the specific patterns of responding across all items. Future studies should examine the utility of CAT with IRT scoring with regard to the GAS.

Despite several strengths of the present manuscript, some noteworthy limitations should be noted. One limitation was the lack of ethnic and educational diversity within the samples. The small number of ethnic minorities prevented ethnic group analyses from being conducted due to limited statistical power (e.g., Cohen, 1992). Future research should examine the psychometric properties of the GAS and GAS-10 in culturally diverse populations of older adults, given the increasing diversity that will be manifested in upcoming cohorts of older adults. Another limitation was the variety of data collection approaches used in the four samples. We combined the groups for some analyses but we recognize the possibility that differences in the data collection strategies may have impacted the results and their generalizability. However, the use of IRT strengthens the generalizability of the results such that the item parameters are generalized beyond the sample used in this study. In addition, data regarding physical health conditions were not available on all study participants, aside from those in Sample 4. Given the high prevalence rates of health conditions in older adults and the high endorsement of somatic symptoms in our study, it is reasonable to assume that some participants from samples 1–3 also had chronic health conditions. Preliminary research (Yochim *et al.*, 2011) has indicated mild-to-moderate correlations between the GAS, subscales, and a measure of medical burden. Future research should explore the potential for DIF by medical conditions and further examine the psychometric properties of the full length GAS and the GAS-10 in diverse medical populations. Furthermore, it would be ideal to understand how the GAS items assess anxiety relative to similar measures with respect to IRT properties. Future studies should continue to utilize IRT to examine the item properties of anxiety measures in older individuals. Finally, the sensitivity and specificity of the GAS in detecting anxiety disorders have not yet been reported. Future studies should do so using the diagnostic criteria for anxiety disorders to determine an appropriate cut-score for clinically significant anxiety. This line of research should also be pursued with the GAS-10 and CAT approaches. Finally, due to the increasing population of older adults across the globe, researchers should pursue culturally sensitive translations of the GAS into other languages.

Overall, this study strongly supports the use of the GAS in assessing anxiety in older adults. As more people throughout the world reach older ages, the number of people affected by this serious condition will increase as a result. As this prevalence increase occurs, empirically supported assessment tools will be increasingly important for use in settings in which older adults receive healthcare. Concurrently, it is necessary to raise healthcare providers' awareness about the prevalence of anxiety in older populations. This study adds to the emerging literature on anxiety assessment in older adults by merging classical test and IRT approaches to assessing psychometrics. Future studies should continue to examine the utility of the GAS in other populations, and continue to merge new and innovative methods for analyzing psychometric properties of assessment measures.

Conflict of interest

None.

Description of authors' roles

Anne E. Mueller designed the study, collected and analyzed data, and wrote the paper. Daniel L. Segal supervised the study, analyzed data, and assisted in writing the paper. Brandon Gavett assisted with data analysis and assisted in editing the manuscript. Meghan A. Marty, Brian Yochim, and Andrea June collected data and assisted in editing the manuscript. Frederick L. Coolidge assisted in editing the manuscript.

Acknowledgment

This paper is based on the doctoral dissertation conducted at the University of Colorado at Colorado Springs by Anne E. Mueller under the supervision of Daniel L. Segal.

References

Baker, F. (2001). *The Basics of Item Response Theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Brock, K., Clemson, L., Cant, R., Ke, L., Cumming, R. G., Kendig, H. and Mathews, M. (2011). Worry in older community-residing adults. *International Journal of Aging and Human Development*, 72, 289–301. doi:10.2190/AG.72.4.a.

Bryant, C., Jackson, H. and Ames, D. (2008). The prevalence of anxiety in older adults: methodological issues and a review of the literature. *Journal of Affective Disorders*, 109, 233–250. doi:10.1016/j.jad.2007.11.008.

Byrne, G. J. and Pachana, N. A. (2011). Development and validation of a short form of the Geriatric Anxiety Inventory – the GAI-SF. *International Psychogeriatrics*, 23, 125–131. doi:10.1017/S1041610210001237.

Calleo, J. *et al.* (2009). Generalized anxiety disorder in older medical patients: diagnostic recognition, mental health management, and service utilization. *Journal of Clinical Psychology in Medical Settings*, 16, 178–185. doi:10.1007/s10880-008-9144-5.

Cairney, J., Corna, L. M., Veldhuizen, B. A., Herrmann, N. and Streiner, D. L. (2008). Comorbid depression and anxiety in later life: patterns of association, subjective well-being, and impairment. *American Journal of Geriatric Psychiatry*, 16, 201–208.

Choi, S. W., Gibbons, L. E. and Crane, P. K. (2011). Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, 39, 1–30.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.

Crane, P. K., Gibbons, L. E., Jolley, L. and van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIF detect and dif with par. *Medical Care*, 44, S115–S123.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: Guilford Press.

De Beurs, E., Beekman, A. T. F., van Dyck, D. J. H. D. and van Tilburg, W. (2000). Predictors of change in anxiety symptoms of older persons: results from the Longitudinal Aging Study, Amsterdam. *Psychological Medicine*, 30, 515–527.

Edelen, M. O. and Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16, 5–18. doi:10.1007/s11136-007-9198-0.

Edelstein, B. A. *et al.* (2008). Older adult psychological assessment: current instrument status and related considerations. *Clinical Gerontologist*, 31, 1–35.

Embretson, S.E. and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Flint, A. J. *et al.* (2010). Effect of age on the frequency of anxiety disorders in major depression with psychotic features. *American Journal of Geriatric Psychiatry*, 18, 404–412.

Gum, A. M., King-Kallimanis, B. and Kohn, R. (2009). Prevalence of mood, anxiety, and substance-abuse disorders for older Americans in the National Comorbidity Survey Replication. *American Journal of Geriatric Psychiatry*, 17, 769–781. doi:10.1097/JGP.0b013e3181ad4f5a.

Kabacoff, R. I., Segal, D. L., Hersen, M. and Van Hasselt, V. B. (1997). Psychometric properties and diagnostic utility of the Beck Anxiety Inventory and the State-Trait Anxiety Inventory with older adult psychiatric outpatients. *Journal of Anxiety Disorders*, 11, 33–47. doi:10.1016/S0887-6185(96)00033-3.

Katon, W., Lin, E. and Kroenke, K. (2007). The association of depression and anxiety with medical symptom burden in patients with chronic medical illness. *General Hospital Psychiatry*, 29, 147–155. doi:10.1016/j.genhosppsych.2006.11.005.

Leach, L. S., Christensen, H. and Mackinnon, A. J. (2008a). Gender differences in the endorsement of symptoms for depression and anxiety: are gender-biased items responsible? *Journal of Nervous and Mental Disease*, 196, 128–135. doi:10.1097/NMD.0b013e318162aa63.

Leach, L. S., Christensen, H., Mackinnon, A. J., Windsor, T. D. and Butterworth, P. (2008b). Gender differences in depression and anxiety across the adult lifespan: the role of psychosocial mediators. *Social Psychiatry and Psychiatric Epidemiology*, 43, 983–998. doi:10.1007/s00127-008-0388-z.

Lowe, P. A. and Reynolds, C. R. (2005). Do relationships exist between age, gender, and education and self-reports of anxiety among older adults? *Individual Differences Research*, 3, 239–259.

Murphy, L. B., Sacks, J. J., Brady, T. J., Hootman, J. M. and Chapman, D. P. (2012). Anxiety is more common than depression among US adults with arthritis. *Anxiety Care & Research*, 64, 968–976. doi:10.1002/acr.21685.

Pedraza, O. and Mungas, D. (2008). Measurement in cross-cultural neuropsychology. *Neuropsychology Review*, 18, 184–193.

Potvin, O. *et al.* (2011). Norms and associated factors of the STAI-Y state anxiety inventory in older adults: results from the PAQUID study. *International Psychogeriatrics*, 23, 869–879. doi:10.1017/S1041610210002358.

R Core Team. (2012). *R: A Language and Environment for Statistical Computing* (computer software). Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.R-project.org; last accessed: 7 December 2012.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34, 100–114.

Schaub, R. T. and Linden, M. (2000). Anxiety and anxiety disorders in the old and very old: results from the Berlin Aging Study (BASE). *Comprehensive Psychiatry*, 41, 48–54. doi:10.1016/S0010-440X(00)80008-5.

Segal, D. L., June, A., Payne, M., Coolidge, F. L. and Yochim, B. (2010). Development and initial validation of a self-report assessment tool for anxiety among older adults: the Geriatric Anxiety Scale. *Journal of Anxiety Disorders*, 24, 709–714. doi:10.1016/j.janxdis.2010.05.002.

Therrien, Z. and Hunsley, J. (2011). Assessment of anxiety in older adults: a systematic review of commonly used measures. *Aging & Mental Health*, 15, 1–16. doi:10.1080/13067863.2011.602960.

Van Dam, N. T., Earleywine, M. and Forsyth, J. P. (2009). Gender bias in the sixteen-item Anxiety Sensitivity Index: an application of polytomous differential item functioning. *Journal of Anxiety Disorders*, 23, 256–259. doi:10.1016/j.janxdis.2008.07.008.

Wolitzky-Taylor, K. B., Castriotta, N., Lenze, E. J., Stanley, M. A. and Craske, M. G. (2010). Anxiety disorders in older adults: a comprehensive review. *Depression and Anxiety*, 27, 190–211. doi:10.1002/da.20653.

Yochim, B. P., Mueller, A. E., June, A. and Segal, D. L. (2011). Psychometric properties of the Geriatric Anxiety Scale: comparison to the Beck Anxiety Inventory and Geriatric Anxiety Inventory. *Clinical Gerontologist*, 34, 21–33. doi:10.1080/07317115.2011.524600.

Yochim, B. P., Mueller, A. E. and Segal, D. L. (2013). Late life anxiety is associated with decreased memory and executive functioning in community dwelling older adults. *Journal of Anxiety Disorders*, 27, 567–575.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.