Adaptation and initial validation of the Patient Health Questionnaire – 9 (PHQ-9) and the Generalized Anxiety Disorder – 7 Questionnaire (GAD-7) in an Arabic speaking Lebanese psy...
Adaptation and initial validation of the Patient Health Questionnaire – 9 (PHQ-9) and the Generalized Anxiety Disorder – 7 Questionnaire (GAD-7) in an Arabic speaking Lebanese psychiatric outpatient sample

Helen Sawaya, Mia Atoui, Aya Hamadeh, Pia Zeinoun, Ziad Nahas*

Department of Psychiatry, American University of Beirut Medical Center, Beirut, Lebanon

1. Introduction

Mood and anxiety disorders are two of the most common and debilitating psychiatric disorders with a lifetime prevalence of 20.8% and 28.8% respectively, in the United States (Kessler et al., 2005). In Lebanon, the lifetime prevalence of mood and anxiety disorders is 12.6% and 16.7%, respectively (Karam et al., 2008). Patients with these disorders often rely on primary health care physicians and physicians in other specialties for diagnosis and treatment (Hijazi et al., 2011), which is sometimes problematic due to the low reliability of primary health care physicians in detecting psychiatric symptoms (El-Rufaie et al., 1997; Becker, 2004; Daradkeh et al., 2005). The availability of quick and reliable screening measures can therefore enhance the reliability and standardization of diagnoses especially when psychiatric services are not readily accessible. Brief screening tools are also useful in psychiatric clinics to help better detect depression and anxiety in patients with comorbid psychiatric illnesses and to prevent misdiagnoses. Two scales¹ that have been widely used as both diagnostic tools and severity measures, are the Patient Health Questionnaire 9 (PHQ-9), and the Generalized Anxiety Disorder 7 (GAD-7).

1.1. The PHQ-9 and GAD-7

The PHQ-9 is a 9-item self-report scale used to assess and monitor depression severity. It was developed as a screening tool for depressive symptomatology but is also used as a diagnostic measure of major depressive disorder (MDD) (Kroenke et al., 2001; Kroenke and Spitzer, 2002). The items on the PHQ-9 correspond to the 9 DSM-IV criteria used to diagnose a major depressive episode (DSM–IV–TR, 2000; American Psychiatric Association, 2000). The scale is shown to have high internal consistency (Chronbach’s alpha between .86 and .95) (Kroenke et al., 2001) and high test-retest reliability (Chronbach’s alpha between .84 and .95) (Kroenke et al., 2001; Löwe et al., 2004). Most studies found evidence for the unidimensionality of the PHQ-9 although two-factor models were also supported (Williams et al., 2009; Huang et al., 2006; Cameron et al., 2011; Krause et al., 2008; Cannon et al., 2007).

Systematic reviews have established the scale’s criterion validity against other measures of depression such as the Hamilton Rating Scale for Depression (HRSD, Cameron et al., 2011), Structured Clinical Interview for DSM disorders (SCID; Thekkumpurath

¹ The PHQ-9 and GAD-7 scales are part of a larger scale consisting of modules measuring mood and anxiety disorders, alcohol abuse, eating and somatoform disorders (Spitzer et al., 1999).
and high specificity subscales were investigated in a sample of Saudi Arabia patients with specific medical conditions (McGuire et al., 2013; Kendel et al., 2010; Navinės et al., 2012; Monahan et al., 2008; Dum et al., 2008) as well as primary care patients and patients in general hospital settings (Daradkeh et al., 2005; Cameron et al., 2008; Persoons et al., 2003; Carballo et al., 2007; Löwe et al., 2004; Eck et al., 2006; Diez-Quevedo et al., 2001). The PHQ-9 has also been translated into different languages and validated in various cultures (example Zhang et al., 2013; Inagaki et al., 2013; Osório et al., 2009; Persoons et al., 2003; Carballo et al., 2007; Navinės et al., 2012; Karekla et al., 2012; Donnelly, 2007; Marc et al., 2014).

The GAD-7 is a 7-item self-report scale, also based on DSM-IV criteria, used to measure the severity of generalized anxiety disorders, and was also shown to be a reliable screening tool for panic, social anxiety and post-traumatic stress disorder (Kroenke et al., 2010; Kroenke, 2007). The GAD-7 exhibits excellent internal consistency (Chronbach’s alpha between .89 and.92) (Löwe et al., 2008; Kertz et al., 2012; Spitzer et al., 2006), good convergent validity when compared to the Beck Anxiety Inventory (r = .72), Symptom Checklist 90 (SCI-90) anxiety subscale (r = .74) (Spitzer et al., 2006), Penn-State Worry Questionnaire (PSWQ-Worry) (r = .64), Depression Anxiety Stress Scale (DASS-anxiety (r = .77)) and DASS-stress (r = .79) (Kertz et al., 2012). Studies also provide evidence for good sensitivity (83–89%) but questionable specificity (46–82%) (Spitzer et al., 2006; Kertz et al., 2012). All studies have shown the underlying structure to be unidimensional. The GAD-7, although not studied as extensively as the PHQ-9, has also been validated in community samples, psychiatric patients and patients in primary health care settings, and has been validated in various cultures (example García-Campayo et al., 2010; Löwe et al., 2008; Delgado et al., 2012; Dear et al., 2011; Donker et al., 2011; Sidik et al., 2012; Konkan et al., 2013).

1.2. Measures of depression and anxiety in Arabic

Despite the wide use of these scales, no study has determined the validity and utility of these scales in the Lebanese population. Other measures of depression and anxiety are available in the Arabic language, but their validity and applicability in Lebanese psychiatric contexts remains uncertain. An Arabic version of the BDI-II was tested on samples of university students from different Arab countries (Alansari, 2005; Alansari, 2006; Al-Musawi, 2001), with results showing good internal consistency but lacking clear measures of validity. The PHQ depression, anxiety and somatization subscales were investigated in a sample of Saudi Arabia primary care patients with results showing high specificity and moderate sensitivity for the depression and somatiform subscales and high specificity but low sensitivity for the GAD and panic subscales, when compared to psychiatric assessment (Becker et al., 2002).

Two depression rating scales that have been validated in Lebanon are the Center for Epidemiologic Studies – Depression Scale (CES-D) (Kazarian and Taher, 2010) and the Hopkins Symptom Checklist – 25 (HSCl-25) (Mahfoud et al., 2013). The CES-D has shown very good internal consistency (alpha = .84) and a significantly higher correlation with a measure of social avoidance (r = .49) compared to social anxiety (r = .26) in a non-psychiatric community sample in Beirut. Factor analysis supported a twofactor solution, with the first factor representing Depressed Affect and the second Lack of Positive Affect. The Depressed Affect factor appeared to be a more accurate and reliable measure of depressive symptoms in the Lebanese population. The psychometric properties of the HSCL-25 were tested against the Mini International Neuropsychiatric Interview on a sample of women with medically unexplained vaginal discharge. Internal consistency was very good for the depression (.88) and anxiety (.85) subscales. Measures of sensitivity were very good for the depression (82%) and anxiety (84%) subscales at an optimal cutoff of 2.10 and 2.00, respectively. Specificity was good for the depression (70%) but fair (59%) for the anxiety subscale.

Although the CES-D and HSCL-25 can be useful screeners for depression and anxiety in the Lebanese population, their length limits their use with difficult or resistant patients and with patients with cognitive impairment. Long scales are also problematic in patient overloaded clinics and busy health care centers. The aim of this study was therefore to adapt and validate the PHQ-9 and GAD-7, two concise and widely used scales in clinical and research settings, as screening tools for depression and anxiety symptoms. The primary aim was to investigate the factor structure and reliability of the Arabic PHQ-9 and GAD-7 in a heterogeneous sample of Lebanese psychiatric outpatients. The second aim was to determine the scales’ sensitivity, specificity and convergent validity by comparing them to clinical diagnoses of depression and anxiety as well as to another self-report measure of depression, the Psychiatric Diagnostic Screening Questionnaire – Major Depressive Disorder (PDSQ-MDD subset. Supplementary material 1).

2. Methodology

2.1. Participants

The study sample included 186 Lebanese adult psychiatric outpatients recruited between 2010 and 2012 during their first visit to the Department of Psychiatry at the American University of Beirut Medical Center (AUBMC). The study was approved by AUBMC’s Institutional Review Board.

2.2. Measures and procedure

The PHQ-9 and GAD-7 were translated from English to Arabic by bilingual psychologists. Two forward translations from English to Arabic were produced for each scale, and then another bilingual psychologist backtranslated both scales into English. The two Arabic versions for each test were compared with their original and back-translated English versions in a series of meetings. Based on consensus, the most appropriate Arabic version was chosen for each test. Minor edits were made to each one. Upon recruitment, participants signed an informed consent and completed a battery of tests comprising a brief demographic form and 4 questionnaires, the PHQ-9, GAD-7, PDSQ and the Inventory of Depressive Symptoms - Self Report (IDS-SR). The order of the questionnaires was counterbalanced across participants. Upon completing the questionnaires, the patients were assessed by one of 4 psychiatrists or 2 clinical psychologists, blinded to the participants’ responses to the questionnaires. The present study only reports on the psychometric properties of the PHQ-9 and GAD-7.

To determine the diagnostic value of the PHQ-9 and GAD-7, participants’ scores on the two questionnaires were compared to the diagnoses made by the clinicians. A total score of 10 was chosen as the cutoff, as this was the most commonly used cutoff and the one that best balances sensitivity and specificity (Kroenke et al., 2001; Spitzer et al., 2006). Any score of 10 or above indicated the presence of a depressive or anxiety disorder and any score of 9 or below indicated its absence. The clinical diagnoses comprising...
depressive disorders were the ones characterized by a depressive component and those comprising anxiety disorders were the ones characterized by an anxiety component (Supplementary material 2). To determine the convergent validity of the PHQ-9, the total scores were correlated with the total scores on the PDSQ-MDD subset. To determine the discriminant validity of the PHQ-9 with the GAD-7, scores on both questionnaires were correlated with scores on the PDSQ-MDD, and the correlations were compared.

2.3. Statistical analysis

The scales’ internal consistency was measured using Cronbach’s alpha estimated from a polychoric correlation matrix. Analysis of the factor structure and dimensionality was done using three methods of exploratory factor analysis (EFA): scree plot, maximum likelihood analysis and parallel analysis, also using polychoric correlation matrices. The criteria used to determine best fit in the maximum likelihood analysis were the cumulative variance explained, the root mean square of the residuals (RMSR) and the communalities. The appropriateness of carrying out factor analysis on the data was determined using measures of sampling adequacy (MSA), Kaiser-Meyer-Olkin (KMO) statistics and Bartlett’s test. To investigate the convergence of the PHQ-9 and GAD-7 scores with clinical diagnoses the scores were used as categorical variables, with 0 and 1 indicating the absence and presence of a depressive or anxiety disorder, respectively. To investigate discriminant validity, Pearson correlations were calculated between the two correlations (PHQ-9 – GAD total score and PDSQ-MDD – GAD total score). Measures of sensitivity, specificity, Cohen’s kappa, and Area Under the Curve (AUC) were also calculated.

3. Results

3.1. Patient characteristics

A total of 186 patients participated in the study. The mean age of the sample was 35.6 and did not differ with respect to gender distribution (p = .34). Mean PHQ-9 (n = 179) and GAD-7 (n = 116) scores were in the moderate symptom severity range. No gender difference in total score on either scale was found (Table 1). Clinician diagnoses were available for 176 of the patients. Fifty-nine percent of the sample (104/176) had a diagnosis of a depressive disorder and 38% (67/176) had a diagnosis of an anxiety disorder. Twenty-six percent (46/176) had comorbid anxiety and depression. The other diagnoses included psychotic disorders, bipolar disorder, personality disorders and others.

Table 1

|                  | Whole sample | Males | Females | Group difference (W|X²|) | p     |
|------------------|--------------|-------|---------|-------------------|-------|
| Age (mean, SD)   | 35.64 (12.98) | 31.87 (12.98) | 37.24 (12.86) | 3297              | 0.04  |
| Gender           | 183          | 85    | 98      | 0.92              | 0.34  |
| PHQ total score  | 12.7 (6.65)  | 12.21 (5.99) | 13.12 (7.17) | 4216              | 0.49  |
| GAD total score  | 10.79 (6.66) | 9.78 (6.66) | 11.81 (6.63) | 1988              | 0.13  |

Note. Age in years; SD: standard deviation; W: Wilcoxon rank sum test with continuity correction; X²: Chi-squared test for given probabilities.

Tables 2 and 3 show the distribution of clinician diagnoses across the sub-categories of depression and anxiety disorders. The distribution of participants’ scores across the PHQ-9 and GAD-7 items were also plotted to determine the items with the highest endorsement of symptoms (Supplementary Figs. 1 and 2). Anhedonia and hopelessness (PHQ-9 items 1 and 2) had the highest percent of participants with severity ratings of 2 and above; nervousness and worry (GAD-7 items 1, 2 and 3) had the highest percent of participants with severity ratings of 2 and above.

3.2. Reliability and factor structure of the PHQ-9

Reliability analyses showed that the items of the PHQ-9 were highly consistent (Chronbach’s alpha = .88). The item-total correlations were high for most items (.62 -.77) and moderately high for two items (>.53). Bartlett’s test was significant, X²(36) = 525.11, p < .0001, and the KMO statistic was high (.83), indicating a good relationship between the data and an adequate sample for factor analysis. The determinant of the correlation matrix indicated the absence of multicollinearity (p = .05). The scree plot indicated a two-factor solution, with the eigenvalue of the first component = 3.9 and the second = 1.2. Factor analysis using maximum likelihood methods was used to test one-, two- and three- factor solutions submitted to oblique rotation (using Oblimin). Comparison of the three models lent support to the three-factor solution (please refer to Supplementary material 3 for details). The output of the two-factor solution showed that items 1, 2, 6, 7, 8 and 9 loaded on factor one and items 3, 4, 5 and 7 loaded on factor two (Table 4). The two factor model explained 54% of the variance, with the first factor explaining 31% of the variance. The two factors correlated moderately (r = .62). The model fit is 98% with the root mean square of the residuals = .06, indicating a good fit. Communalities were adequately high on all items. Parallel analysis was also carried out to confirm the factor structure of the data. The output suggests that the number of factors to be extracted is three. Due to the difficulty of interpreting the three-factor solution, we decided that the two-factor model is more
Table 4
Factor loadings of the PHQ-9 items in the two-factor solution.

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1 (mood)</th>
<th>Factor 2 (somatization)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Little interest or pleasure in doing things</td>
<td>0.69</td>
<td>0.12</td>
</tr>
<tr>
<td>2. Feeling down, depressed or hopeless</td>
<td>0.8</td>
<td>0.04</td>
</tr>
<tr>
<td>3. Trouble falling asleep, staying asleep or sleeping too much</td>
<td>-0.03</td>
<td>0.71</td>
</tr>
<tr>
<td>4. Feeling tired of having little energy</td>
<td>0.13</td>
<td>0.72</td>
</tr>
<tr>
<td>5. Poor appetite or overeating</td>
<td>0.1</td>
<td>0.74</td>
</tr>
<tr>
<td>6. Feeling bad about yourself, that you’re a failure or have let yourself or family down</td>
<td>0.7</td>
<td>-0.01</td>
</tr>
<tr>
<td>7. Trouble concentrating on things</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>8. Moving or speaking slowly, or being fidgety or restless</td>
<td>0.53</td>
<td>0.22</td>
</tr>
<tr>
<td>9. Thoughts of suicide or hurting yourself</td>
<td>0.79</td>
<td>-0.13</td>
</tr>
</tbody>
</table>

Table 5
Factor loadings of the GAD-7 items in the uni-dimensional model.

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1 (anxiety)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Feeling nervous, anxious or on edge</td>
<td>0.91</td>
</tr>
<tr>
<td>2. Not being able to sleep or control worrying</td>
<td>0.93</td>
</tr>
<tr>
<td>3. Worrying too much about different things</td>
<td>0.95</td>
</tr>
<tr>
<td>4. Trouble relaxing</td>
<td>0.83</td>
</tr>
<tr>
<td>5. Being so restless that it’s hard to sit still</td>
<td>0.67</td>
</tr>
<tr>
<td>6. Becoming easily annoyed or irritable</td>
<td>0.86</td>
</tr>
<tr>
<td>7. Feeling afraid as if something awful might happen</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Further analyses were carried out in an attempt to understand reasons behind the low specificity of the PHQ-9 by examining the cases in which a clinical diagnosis other than depression was made for those who scored above cutoff on the PHQ-9 (false positive cases on the PHQ-9). Clinician diagnoses for those participants were the following: OCD (n=10), panic with or without agoraphobia (n=5), GAD (n=3), anxiety NOS (n=4), adjustment disorder (n=4), somatization (n=1), marital discord (n=1), ADD (n=1), sleep disorder (n=2), sexual disorder (n=2), psychosis (n=1), opiate dependence (n=1), Tourette’s syndrome (n=1), no diagnosis (n=2). (Note: 1 participant with panic had comorbid OCD).

3.5. Sensitivity and specificity of the GAD-7

Based on a cutoff of 10, the GAD-7 showed low sensitivity (57%) and low specificity (53%) when compared to clinician diagnoses. Measure of Cohen’s kappa was 0.1 indicating very low agreement between the scale scores and clinical diagnoses. AUC for this sample was 0.57 (Supplementary Fig. 4), indicating that the GAD-7 has no discriminatory power. The other cutoffs (total scores of 8, 9, 11 and 12) did not show better psychometric properties of the GAD-7 as the sensitivity and specificity values were also low. Further investigation was carried out to understand reasons behind the inconsistencies. Patients who scored 10 or above on the GAD-7 and were not given a clinical diagnosis of anxiety (false positive cases) had the following diagnoses: MDD, recurrent or chronic MDD (n=11), bipolar I or II (n=4), dysthymia (n=2), adjustment disorder (n=2), OCD (n=6), substance abuse related disorder (n=2), somatization (n=1), no diagnosis (n=2). (Note: 1 participant with OCD had comorbid depression).

3.6. Correlation between the PHQ-9 and GAD-7 scales

Pearson correlation revealed a large positive correlation (r=0.72, p<0.000) between the scores on the two scales. To rule out the possibility that the strong correlation was driven by the participants with comorbid depression and anxiety, the analysis was repeated excluding those participants. Results reveal a similar finding (r=0.77, p<0.000).

3.7. Convergent and discriminant validity of the PHQ-9 and GAD-7

Convergent validity was calculated by correlating total scores on the PHQ-9 with total scores on the PDQSQ-MDD subset (r=.75). Discriminant validity was calculated using Steiger’s z by comparing correlations (n=107) of the PHQ-9-PDSQ-MDD (r=.76) and the GAD-7-PDSQ-MDD (r=.51), z=4.75, p<0.00001. This result reveals that the PHQ-9 is significantly more related to another measure of depression than the GAD-7 is.

Another way to determine whether the items on the PHQ-9 and GAD-7 scales tap into different constructs, is by running a joint factor analysis on the items of both scales to determine the pattern of factor loadings. Results of the two-factor solution showed a separation between items of the PHQ-9 and GAD-7 scales, with the exception of PHQ-9 item 1 (anhedonia) and item 2 (feeling down), which loaded on both factors 1 and 2 (Table 6).

4. Discussion

The aim of this study was to translate and provide psychometric properties for two quick screening measures of depression and anxiety to be used in medical settings around Lebanon for clinical and research purposes. Based on the distribution of scores across the scale items results show that participants endorsed...
more severe PHQ-9 symptoms of anhedonia and hopelessness compared to other symptoms, and more severe GAD-7 symptoms of nervousness and worry. Results support the reliability of the PHQ-9 and GAD-7 and their factor structures replicate those found in other studies. Findings related to the convergence of results between the PHQ-9 scores and independent clinical diagnoses were mixed. There was evidence for good sensitivity but poor specificity when compared to clinician diagnoses, good convergent validity when compared to the PDSQ-MDD subset and good discriminant validity with the GAD-7. The convergence of GAD-7 scores with independent clinical diagnoses was poor. The poor convergence on anxiety was also noted in a study comparing clinical diagnoses to those derived from a semi-structured interview. Authors found the least agreement to be on anxiety, as opposed to agreement on mood and disruptive disorders in adolescents (Zeinoun et al., 2013).

Meta-analyses of PHQ-9 psychometric properties showed good sensitivity (77%, 80%) and excellent specificity (94%, 92%), although there is evidence of substantial heterogeneity in validity measures across studies (Wittkampf et al., 2007; Gilbody et al., 2007). In an Ethiopian sample of general hospital outpatients, the PHQ-9 showed very good sensitivity (86%) and moderate specificity (67%) (Gelaye et al., 2013). In a Swiss sample of internal medicine patients, the PHQ-9 had low sensitivity (50%) in detecting MDD and moderate sensitivity (65%) in detecting all depressive disorders (Carballeira et al., 2007). In a Japanese psychiatric sample, sensitivity was excellent (94%) but specificity was poor (50%) (Inoue et al., 2012). In a Malay primary care sample, the PHQ-9 showed moderate sensitivity (69%) and low specificity (61%) with an optimal cutoff of 5 (Azah et al., 2005).

The varying accuracy of the PHQ-9 across different studies could be explained by differences in language, patient population or the gold standard measure used for comparison. Few very studies have tested the validity of the PHQ-9 in a psychiatric sample. One study that did (Inoue et al., 2012), also showed evidence for low specificity (50%). Low specificity of this measure in psychiatric samples is not surprising given that depressive symptoms are prevalent in most psychiatric disorders. The PHQ-9 therefore picks up on these symptoms even in patients with psychiatric disorders other than depression. The depressive symptoms detected by the PHQ-9 might therefore not be considered by the clinicians to fit a separate diagnostic category of depression.

Another explanation for the low specificity of the PHQ-9 is that the scale itself tends to overestimate depressive symptoms. There is evidence that the PHQ-9 categorizes participants in higher categories representing more severe symptomatology compared to other measures of depression. In a UK sample of primary health care patients, the PHQ-9 categorized to a larger extent participants in the severe category compared to the HRSD-17 and HAD-D (Cameron et al., 2011; Cameron et al., 2008). In a US sample of patients with substance abuse disorder, the PHQ-9 categorized to a larger extent individuals in the mild category compared to the BDI-II, which categorized these individuals as having no depressive symptoms (Hepner et al., 2009). Another study measuring the psychometric properties of three depression scales in a sample of patients with Hepatitis C showed that although the CES-D and BDI-II had moderate agreement (kappa = .64), the PHQ-9 had low agreement with both scales (Kappa = .18 and .24, respectively), most likely due to the PHQ-9’s over-detection of depressive symptoms (Dbouk et al., 2008). Results from these studies could indicate that the PHQ-9 exaggerates depressive symptomatology or that it is more sensitive to sub-threshold depressive symptoms.

The poor psychometric properties of the GAD-7 in the present study contrasts with studies showing good sensitivity and specificity of the scale, such as in healthy and psychiatric Spanish samples (86% and 93%, respectively) (García-Campayo et al., 2010), a Malay sample of women in primary care (76% and 94%) (Sidik et al., 2012) and a Turkish sample of psychiatric outpatients (80% on both measures at an optimal cutoff of 8) (Konkan et al., 2013). Other reports, however, have shown moderate and low specificities for the GAD-7, such as in a Dutch sample from the general population (65% at a cutoff of 12) (Donker et al., 2011) and in a US psychiatric sample (46% at a cutoff of 10) (Kertz et al., 2012). A few explanations for the low specificity of the GAD-7 have been proposed. Some authors suggest that the scale measures negative affect more broadly rather than anxiety specifically (Donker et al., 2011; Kertz et al., 2012), thus over-representing anxiety in both healthy and patient populations. Others suggest that the GAD-7 is not exclusive to measuring anxiety since it consists of items that tap into aspects of depression (García-Campayo et al., 2010). There is evidence that the GAD-7 correlates strongly with measures of depression (DASS – depression and CES-D-10 depression) (Kertz et al., 2012).

Low specificity of the GAD-7 is not a problem specific to this scale, however, since other measures of anxiety have also shown low discriminant validity with depression. The STAI-T and Beck Anxiety Inventory (BAI) have shown stronger correlations with measures of depression than measures of anxiety (Balsamo et al., 2013; Fydrich et al., 1992; Wetherell and Gatz, 2005). Based on these findings researchers have created factor analytic models in an attempt to understand the relationship between depression and anxiety constructs. Although some studies have provided evidence for the distinctness of the two constructs (Spitzer et al., 2006), others have shown that they share more common than unique variance (Feldman, 1993). The most agreed upon models posit that the core depression and anxiety constructs, characterized by unique symptoms each, are distinct, but correlate strongly with a higher-order factor that consists of non-specific symptoms common to both disorders (example poor concentration, sleep difficulties and psychomotor agitation) (Watson et al., 1995; Kovacs et al., 2004). This could explain why the two scales in the present study were highly correlated even in patients without comorbid depression and anxiety. The PHQ-9 and GAD-7 consist of numerous items that are not specific to depression or anxiety. The two scales even contain an item that measures the same symptom (restlessness), which has been shown in another study (Kertz et al., 2012) to have the lowest item-total correlation, indicating that it is not specific to anxiety.

The overlap between depression and anxiety constructs appears even more pronounced in Arabic cultures. A factor analytic study on the HSCL-25 in Arab college students in Kuwait showed that the variances in both depression and anxiety subscales were
largely explained by a general “negative affect” factor (Al-Turkait et al., 2011). A study on the Depression Anxiety Stress Scales (DASS) that aimed at comparing the English and Arabic versions of the questionnaire, showed that although the DASS adequately discriminated between the three constructs (depression, anxiety and stress), there was less differentiation in the Arabic version compared to the English one (Taouk et al., 2001). A study done on children and adolescents in Lebanon using the Screen for Child Anxiety Related Disorders also showed that the anxiety scale did not discriminate between depressive and anxiety disorders (Hariz et al., 2013). The depressed participants in this study scored highly on the scale, possibly pointing to the failure of the children to differentiate between their own symptoms of depression and the anxiety symptoms described by the questionnaire items. In line with these findings, results of the present study showed that a large number of patients with a depressive disorder scored highly on the GAD-7.

The large overlap between the two constructs is further evidenced by the finding that when items on both PHQ-9 and GAD-7 scales were combined into a single factor analysis, PHQ-9 items ‘anhedonia’ and ‘feeling down’, the two most central symptoms of depression based on the DSM-IV, loaded on the depression factor as well as the anxiety factor. This finding suggests that anxiety in our sample was also conceptualized in terms of feeling down and taking little interest in doing things. This contrasts with a meta-analysis that showed separate loadings for the PHQ-8 and GAD-7 items when run in a single factor analytic model (Kroenke et al., 2010).

Results from studies done with Arabic-speaking participants could indicate that depression and anxiety disorders are characterized by more similar symptoms in Arab countries (possibly somatic symptoms). It has been shown that Arab populations express their distress using both psychological and somatic descriptors without necessarily differentiating the two (Kazarian and Taher, 2010; Ghushash et al., 2000; Beshai et al., 2013), which could suggest a more subtle discernment between symptoms characterizing different disorders. This pattern was not seen in the present study, however, where participants clearly differentiated mood and somatic symptoms. One explanation for the discrepancy could be the different questionnaires and populations tested, with the studies described above using the CES-D on community and student samples. Our two-factor solution on the PHQ-9 and the higher endorsement of mood and cognitive symptoms (guilt, self-criticism and loss of pleasure) as opposed to somatic symptoms, are evidence against the idea that our participants had a somatic presentation of their disorder. It is therefore improbable that the large overlap between depressive and anxiety symptoms in our sample can be explained by somatization. Future studies using confirmatory factor analysis are needed to determine whether the mood and somatic components measured by the PHQ-9 are indeed separate in the Lebanese culture.

5. Limitations

There are several limitations to the present study. The main limitation is that scale scores were compared to clinical diagnoses. Our clinicians do not have a common gold standard tool to make their diagnoses, such as a validated structured interview in Arabic. Diagnoses were therefore collected through chart reviews. Moreover, clinicians consisted of psychiatrists and clinical psychologists, with varying degrees of seniority. This may have also resulted in diagnostic variation, especially in the degree to which comorbidities were noted. Visual inspection of the clinician diagnoses, however, did not show a difference in the number of depression or anxiety diagnoses made by each clinician. Another important issue to consider is the possibility of differential functioning of the scales in our population. There is evidence that the same scales in different cultures have different factor structures (example Reda, 2011), optimal cutoffs (Mahfoud et al., 2013; El-Rufaie and Absood, 1995) and symptom endorsement (Alansari, 2005). There is scarcity of methodologies available for ensuring linguistic and cultural relevance in diglossic languages like Arabic. Although there are several guidelines proposed for translation of instruments from one language to another (e.g. Geisinger 1994), there is limited literature on methodology for translating into Arabic – a diglossic language whereby the spoken language is qualitatively different from the written language (Ryding, 1991). The technique of translation and back translation that we used, albeit widely used and cost-efficient, may have not been sufficient to correctly capture the meaning of the items, hence introducing bias at the item level (van de Vijver and Poortinga, 2005). When adapting scales one also needs to consider the response options (Maters et al., 2013) and differences in the manifestation of symptoms and conceptualization of the disorders (El-Rufaie and Absood, 1995; Mahfoud et al., 2011). Future studies can develop more robust methodologies of translation and adaptation to establish equivalence of instruments at the level of construct and scores (van de Vijver and Poortinga, 2005). Confirmatory factor analysis would also be needed to confirm the factor structure of the data.

A third limitation is the validation of screening measures in a clinical population, where the prevalence of depressive and anxiety disorders is high and comorbidity is the norm rather than the exception. Specificity of the PHQ-9 might be higher if measured in a sample from the general population. Moreover, although convergent and discriminant validity of the scales were established, the PDSQ is not a validated tool in Arabic and its psychometric properties in this population are unknown.

In conclusion results of the present study show that the PHQ-9 is a sensitive measure of depressive symptoms in a psychiatric sample of Lebanese adults. It can therefore be a useful screening tool for depression not only in psychiatric clinics but also in settings that lack sufficient psychiatric care to determine the need for psychiatry referral. Although its low specificity prevents it from being an adequate tool to rule out depression, its discriminant validity was established, showing that depression and anxiety are distinguishable, albeit correlated. The GAD-7 showed low sensitivity and specificity indicating that in its present form it might not be an adequate tool to assess anxiety symptoms. Future studies are needed to further investigate the reasons behind the poor psychometric properties of the GAD-7. A bottom up approach needs to be considered to better understand how the Lebanese population conceptualizes, understands and expresses symptoms of depression and anxiety.

Acknowledgment

Work was funded by PFIZER WI74077 (ZN).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.psychres.2016.03.030.
1679–1686.