Health and Primary Care

Research Article ISSN: 2515-107X

Psychometric analysis of the Generalized Anxiety Disorder Scale and the Patient Health Questionnaire using Mokken scaling and confirmatory factor analysis

Luke Boothroyd[1], David Dagnan[2] and Steven Muncer[1]*

[1]Clinical Psychology, Psychology Department, University of Teesside, Middlesbrough, UK

[2]Cumbria NHS, UK

Abstract

The Generalized Anxiety Disorder Scale (GAD-7) and Patient Health Questionnaire Depression Scale (PHQ-9) are widely used measures in primary health care used to assess anxiety and depression. While there is no doubt about their reliability and validity, there is some question over their factor structure. Many have suggested that both are one factor scales. However, there is some evidence that they can also be seen as having two subscales that correspond to a Cognitive/Affective aspect and a Somatic aspect. In this study the dimensional structure of both scales is examined using Mokken analysis which is an Item Response Theory approach, and also confirmatory factor analysis. The relationship of the scales two each other and also to the Work and Social Adjustment Scale (WSAS) is also investigated. There is evidence to support the idea of a Cognitive/Affective and Somatic subscale in both measures. While both measures are positively correlated with the WSAS, the PhQ-9 and indeed the PHQ-2, have significantly stronger relationships with it than the GAD-7.

Introduction

Generalized anxiety disorder and depression are among the most frequent disorders in primary care with a prevalence rate of about 8 – 10 % [1], and also show a high degree of comorbidity [2]. The seven item Generalized Anxiety Disorder Scale (GAD-7) [3]is a widely used measure of anxiety which has been shown to have good psychometric properties on numerous occasions. Most recently Jordan, *et al.* [4] investigated the properties of the GAD 7 using item response theory as well as classical test theory methods. Similarly, the Patient Health Questionnaire Depression Scale (PHQ9) [5,6] has also been shown to have good psychometric properties. Ryan, *et al.* [7] for example, have shown that the factor structure of the PHQ 9 was not affected by different methods of data collection, face to face or telephone interview. They are both, therefore, widely used tests with a fair amount of evidence attesting both to their reliability and validity.

There are, however, still some areas which are open to discussion or have contradictory positions. For example, there is disagreement over the factor structure of the PHQ-9. Many authors have suggested that the scale is best seen as a one factor scale. Ryan, *et al.* [7] for example found that a one factor model provided a good fit in his sample of 23672 patients from the UK's Improving Access to Psychological Therapies (IAPT) programme, as long as some of the error covariances were allowed to correlate. The PHQ-9, however, has not always been found to fit a single factor model. Beard, *et al.* [8] studied 1,023 psychiatric participants who completed the PHQ-9 at admission and discharge from an outpatient programme. Confirmatory factor analysis (CFA) suggested a two-factor solution; the first factor represented cognitive and affective symptoms whilst the second factor reflected somatic symptoms. Furthermore, Elhai, *et al.* [9] study of 2,615 Army National Guard Soldiers in Ohio, USA used CFA to evaluate three, two-factor models previously established in the literature. A two-factor model (X^2 = 210.35, p <0.001, CFI = .96, TLI = .94, RMSEA = .05) fitted the data better than a single factor model (X^2 = 317.71, p <0.001, CFI = .94, TLI = .91, RMSEA = .06). The preferred two-factor model reflected a somatic factor and a cognitive-affective factor of depressive symptoms. The cognitive-affective items loading on to factor 1 were items 1 (Anhedonia), 2 (Depressed mood), 6 (Feelings of worthlessness) and 9 (Suicidal ideation). Items 3 (Sleep difficulties), 4 (Fatigue), 5 (Appetite changes), 7 (Concentration difficulties) and 8 (Psychomotor agitation) loaded on to the somatic.

Whilst Spitzer, *et al.* [3] criterion standard study has been supported by several studies in different populations, the GAD-7 has also been found to have a different factor structure to that discovered originally [10,11]. Within an acute psychiatric population (N = 232) in the US Kertz, *et al.* [10] found that although the GAD-7 showed excellent internal consistency (Cronbach's α = 0.91), confirmatory factor analysis failed to support a unidimensional factor structure. The sample included patients with a diagnosis of: social anxiety disorder (n = 42), panic disorder (n = 27), obsessive-compulsive disorder (OCD, n = 25) and PTSD (n = 19). Kertz, *et al.* [10] found that items 5 ('Being so restless that it is hard to sit still') and 6 ('Becoming easily annoyed or irritable') loaded only moderately (0.52 and 0.53 respectively) on to the latent factor in comparison to all other items (0.64 - 0.81). A unidimensional factor structure was only found to be a good fit if items 4 ('Trouble relaxing'), 5 and 6 could co-vary. Whilst the sample of each anxiety disorder was relatively small, it is suggested that the GAD-7 may perform differently in anxiety disorders other than GAD.

****Correspondence to:*** Steven Muncer, Clinical Psychology, Psychology Department, University of Teesside, Middlesbrough, TS1 3BA, UK, E-mail: S.Muncer@tees.ac.uk

Received: July 16, 2018; **Accepted:** July 24, 2018; **Published:** July 27, 2018

A larger scale study of patients receiving brief intensive CBT at a partial hospital program (N = 1,082) in the US by Beard and Bjorgvinsson [11] found the GAD-7 to have psychometric properties like those found in Kertz et al. [10] study. Of the 1,082 patients, 108 (11.7%) had a primary diagnosis of panic disorder, 96 (10.4%) had a primary diagnosis of PTSD and 89 (9.8%) had a primary diagnosis of OCD. The GAD-7 demonstrated good internal consistency across the total sample (Cronbach's α = 0.88). A rotated 2-factor structure was found to account for 70% of the variance. Within the 2-factor structure, the first factor included items 1 (Feeling nervous, anxious or on edge), 2 (Not being able to stop or control worrying), 3 (Worrying too much about different things) and 7 (Feeling afraid as if something awful might happen). The second factor included the remaining items 4, 5 and 6. This 2-factor structure supports the findings of Kertz, *et al.* [10] This could suggest a separate cognitive-affective and a somatic or behavior factor measured by the GAD-7, which has also been highlighted in studies of the PHQ-9.

The GAD-7 was found to have a single factor structure in predominantly primary care samples. This was not supported in a psychiatric sample and samples that included a range of anxiety disorders. A two-factor structure was found which appeared to separate GAD-7 items reflecting the cognitive and emotional experiences of anxiety (items 1, 2, 3 and 7) from items that reflected more physical, behavioural manifestations of anxiety (4, 5 and 6).

In 2009 Kroenke, *et al.* [6] provided evidence for an ultra-brief screening scale called the PHQ-4, which was based on a combination of the PHQ-2 ('Feeling down, depressed or hopeless' and 'Little interest or pleasure in doing things') and GAD-2 (Feeling, nervous anxious or on edge' and 'Not being able to stop or control worrying'), taking the first two items from each of the scales which have already been shown to be good for screening. Interestingly the analysis of the GAD-7 by Jordan, *et al.* [4] also found that the first item pair was better than almost all the others with the possible exception of the second and third item pair, but so far there has been no item response theory analysis of the PHQ-9.

Another area of interest with these scales is in the possible overlap of symptoms and items. There is clearly often comorbidity between depression and anxiety with the two co-occurring at as much as 50% of the time [12]. The strong correlation between the two scales is both evidence of convergent validity in reflecting the comorbidity of the disorders, but also might indicate the possibility of redundancy among items. In general, psychometric analysis has treated the two scales as separate and conducted two sets of analyses on the items. It would be useful to analyze the items as if they were one scale and see if there is redundancy, and also investigate the number of factors needed to explain all 16 items.

In the current study an item response theory approach, Mokken analysis, will be adopted to investigate the items and separate scales, in a similar way to which Jordan, *et al.* [4] investigated the GAD-7. Mokken scaling is a non-parametric method of item response theory which can be used to investigate the dimensional structure of scales. Mokken scaling is similar to Rasch scaling techniques but has the advantage of having fewer restrictions in its use [13]. Although based on Guttman scaling, Mokken does not assume error-free data. Nor does it include assumptions about the sigmoid shape of item characteristic curves that can result in the rejection of many items and so decrease the reliability of the resultant measure. Confirmatory factor analysis will also be used to investigate the factor structure of the items both as separate scales and together. Lastly, we will look at the relationship of the various scales and subscales to a simple measure of impairment in functioning, the Work and Social Adjustment Scale [14].

Method

Participants

Questionnaire data from seven thousand seven hundred and sixty-three patients (38% male; 62% female) registered with an IAPT service in the North of England were examined. The data was collected between February 2009 and August 2015.

Data Analysis

Cronbach's alpha and the Molenaar Sijtsma (MS) statistic were calculated as measures of reliabilitiy. Confirmatory factor analysis (CFA) was carried out using Lavvan package in R [15]. Diagonally weighted least squares estimation with correction to means and variances was used as it is considered to be the best estimator for categorical data. The Comparative Fit Index (CFI) and Root Mean Square Error of Approximation were used as measures of fit. Mokken analysis was used to further understand the structure of the scales using the Mokken package in R [16]. Loevinger's coefficient (H) is the most important calculation in Mokken scale analysis. The basis of Loevinger's coefficient is the extent to which pairs of items conform to Guttman criteria. Scores on pairs of items should consistently be relative to one another. That is, an item that is more or less likely to be endorsed than another should be consistently so across participants. The 'difficulty' of an item refers to how easily an item of a scale is agreed with by respondents; more difficult items have lower mean scores. If the easier to endorse item is endorsed less than the more difficult item then this is a Guttman error. In this case for a PHQ-9 item, a higher depression level should lead to a higher score on the item. Loevinger's H calculates the size of this error for each item, pairs of items and the overall scale. H values of 0.5 indicate a strong scale; weak scales are represented by H values of 0.4 and below. The automated item selection procedure in Mokken was used with the default scalability criterion of H > 0.3 for each item. This procedure is used to maximize the overall H value of the scale.

Results

Mokken Analysis

Cronbach's alpha for the sixteen items as a scale was 0.91 and the MS statistic was also 0.91, suggesting the scale has good reliability. The individual item H values are presented in Table 1. The items appear in the table in the order that they entered into the scale. As can be seen six of the GAD-7 items were first added to the scale, with items 2 and 3 ("Not being able to stop worrying" and "Worrying too much") coming out first. The nine items from the PHQ-9 are added to the scale next, with each item having an H value above 0.36, and above the criterion of .3. As the items are added to the scale the H value of the scale decreases from 0.79 to 0.43 when all items are included. It is interesting to note that the "Become easily annoyed." item from GAD-7 is the last item to enter. A scale which consisted of all sixteen items would have an H value of 0.43 which suggest that this is only a moderate scale (Table 1).

The scales were then considered separately. For the GAD-7 the order of entry was the same with "Become easily annoyed" as the last item entered. The pattern of results ($r_s = 0.87$) is similar to that found by Jordan, *et al.* [4] and supports their contention that using item 2 and item 3 as an alternative to items 1 and 2 in a two-item version of the GAD7 may be possible. On this occasion the H value of the scale is 0.52,

Table 1. Results from Mokken analysis of GAD7 and PHQ9

GAD 7 and PHQ 9 items together	Item H	GAD 7 and PHQ items separately	Item H	GAD7 and PHQ items as four scales	Item H
Not being able to stop worrying…	0.48	*Stop worrying…*	0.59	*Stop worrying…* Cognitive/Affect	0.68
Worrying about different things…	0.48	*Worrying about things…*	0.59	*Worrying about things…* Cognitive/Affect	0.65
Feeling nervous anxious…	0.46	*Feeling nervous anxious…*	0.55	*Feeling nervous anxious…* Cognitive/Affect	0.65
Trouble relaxing…	0.48	*Trouble relaxing…*	0.56	*Feeling afraid…awful…..* Cognitive/Affect	0.55
Being so restless…	0.42	*Being so restless…*	0.49	*Being so restless…* Somatic	0.54
Feeling afraid as if something awful…..	0.37	*Feeling afraid…awful…..*	0.48	*Trouble relaxing…* Somatic	0.54
Moving or speaking slowly or fidgety…..	0.42	*Becoming easily annoyed…*	0.41	*Becoming easily annoyed…* Somatic	
Trouble concentrating on things….	0.45	Feeling down, depressed	0.54	Feeling down, depressed……..Cognitive/Affect	0.65
Feeling down, depressed or hopeless	0.49	Little interest or pleasure	0.53	Little interest or pleasure……..Cognitive/Affect	0.58
Little interest or pleasure in doing…..	0.46	Feeling bad about yourself .	0.47	Feeling bad about yourself …...Cognitive/Affect	0.58
Feeling bad about yourself or that…..	0.44	Thoughts that better off dead	0.43	Thoughts that better off dead…Cognitive/Affect	0.56
Thoughts that better off dead…..	0.39	Trouble concentrating…	0.48	Feeling tired…… Somatic	0.48
Feeling tired……	0.41	Feeling tired……	0.48	Trouble falling or staying asleep… Somatic	0.47
Trouble falling or staying asleep…	0.39	Trouble falling or staying asleep…	0.45	Poor appetite or overeating Somatic	0.44
Poor appetite or overeating	0.38	Poor appetite or overeating	0.44	Trouble concentrating… Somatic	0.46
Becoming easily annoyed or irritable	0.37	Moving or speaking slowly or	0.41	Moving or speaking slowly or Somatic	0.47
Scale	0.43	Scale (GAD 7; PHQ 9)	0.52; 0.47	Scale (GAD/Cog; GAD/Somatic; PHQ/Cog;PHQ/Somatic	0.62;0.50 0.59;0.46

suggesting a strong scale. The Cronbach's alpha of 0.86 and MS or 0.87 are also acceptable values.

Not surprisingly the order of item entry into the scale is different for the PHQ9 when analyzed separately. Here "Feeling down, depressed or hopeless" and "Little interest or pleasure in doing things" are the first two items, and these are the two items that Kroenke, *et al.* [6] suggest for the PHQ4. The scale overall has a lower H value of 0.47 which puts it into the moderate category. Overall the results from the Mokken analysis at this stage suggest that it is best to see the GAD-7 and PHQ-9 as two scales measuring related but different constructs.

Mokken analysis was also conducted on the four possible subscales that have been suggested for the PHQ-9 and GAD-7. In both cases these can be seen as somatic/behavioural and cognitive/affective. The Cognitive PHQ-9 subscale consists of four items and has good reliability statistics with a Cronbach's alpha of 0.8 and an MS of 0.81. All four of the proposed items enter the scale with the lowest item H value of 0.56 for 'Thoughts that you would be better off dead". The overall subscale has an H value of 0.59 which would make if a strong scale. The Somatic PHQ-9 subscale's five items have a Cronbach's alpha of 0.77 and an MS of 0.78. All of the proposed items enter the scale with the lowest H item value of 0.43 for "Moving or speaking so slowly…". The overall subscale H value was 0.46 which would make it a moderate scale.

The Cognitive GAD-7 subscale consists of four items and has a Cronbach's alpha of 0.84 and an MS of 0.85. All of the items entered into the scale and the lowest H item value of 0.55 for "Feeling afraid as if something awful." The overall subscale H value of 0.62 suggested a very strong scale. The Somatic GAD-7 subscale has a Cronbach's alpha of 0.72 and an MS of O.72. All three of the proposed items enter the scale with the lowest item H value of 0.42 for "Becoming easily annoyed…". The overall subscale has an H value of 0.503 which would be a strong scale.

Confirmatory Factor Analysis

The results of the confirmatory factor analysis are presented in Table 2. Not surprisingly given previous results and the Mokken analysis, a one factor solution for the combined 16 items from the GAD-7 and PHQ-9 is not a good fit. Both the GAD-7 and PHQ-9 show reasonable fit to a one factor model when considered separately, but both are significantly better represented by a two factor model with items identified as Cognitive/Affective and Somatic ($\Delta X^2 = 833$, $p < .001$; $\Delta X^2 = 660$, $p < .001$). A model with four factors representing each of the subscales is a good fit to the data with an RMSEA of 0.06 and a CFA of 0.99, which is again significantly better than the one factor model (ΔX^2 (6) = 9240, $p < .001$). Overall the CFA supports the results of the Mokken analysis in suggesting that while a one factor solution for each of the scales is reasonable, two factors provide better fit and scale statistics (Table 2).

Relationship with The Work and Social Adjustment Scale (WSAS)

The WSAS has a Cronbach's alpha and an MS of 0.79. The correlation between the various scales and subscales of the GAD-7 and PHQ-9 are presented in Table 3. All of the scales and subscales of the PHQ-9 and GAD-7 correlate significantly with the WSAS. The PHQ-9 is a significantly better predictor of WSAS total score than the GAD-7 ($t(7760) = 15.36$, $p < .001$). Interestingly the PHQ-4 ($t(7760) = 11.96$, $p < .001$) and also the PHQ-2 which consists of the first two items in the PHQ9, are also significantly ($r = 0.50$; $t(7760) = 6.54$, $p < 0.01$)) more correlated with WSAS scores than the GAD7 (Table 3).

Discussion

The pattern of results suggests that although the items from the GAD-7 and PHQ-9 can be considered as one scale; from a Mokken point of view this would be a weak scale, and also with poor fit from a more traditional classical test theory standpoint. There is much stronger evidence that they should be considered as two separate scales from both the Mokken and confirmatory factor analysis. It should be noted, however, that the PHQ-9 is not as strong as the GAD-7 from a Mokken standpoint. There is also good evidence that within the two scales, it is possible to find two subscales and three of these four subscales would be regarded as strong scales. The results from our Mokken analysis of the GAD-7 show similarity to those of Jordan, *et al.* [4] except that we carried the analysis one stage further by examining the possibility of subscales. It should be noted that the subscales do not appear when straightforward items selection procedures are used.

Table 2. Fit statistics for models of GAD7 and PHQ9

Model	X2/Df	RMSEA and 90%CI		CFI
One factor 16 items	12263/104	0.123	0.121-0.125	0.965
One factor GAD7	1172/14	0.103	0.098-0.108	0.992
Two factor GAD7	239/13	0.047	0.042-0.053	0.998
One factor PHQ9	1479/27	0.083	0.080-0.087	0.986
Two factor PHQ9	819/26	0.063	0.059-0.066	0.993
Four factors	3023/98	0.062	0.060-0.064	0.992

Table 3. Correlation between GAD7, PHQ9 and WSAS

Scale	GAD7	GADCog	GADSom	PHQ9	PHQCog	PHQSom	PHQ4	WSAS
GAD + PHQ (16 Items)	0.89	0.81	0.81	0.93	0.83	0.87	0.90	0.56
GAD7		0.93	0.88	0.65	0.58	0.61	0.82	0.44
GAD Cognitive			0.64	0.57	0.54	0.51	0.82	0.38
GAD Somatic				0.62	0.52	0.61	0.64	0.43
PHQ9					0.89	0.93	0.81	0.56
PHQ Cognitive						0.68	0.83	0.52
PHQ Somatic							0.68	0.51
PHQ4								0.51

All of the subscales are significantly positively correlated with WSAS. There is a significant difference in the strength of the correlation, with the PHQ-9 being significantly more correlated. Perhaps even more importantly both of the PHQ-9 subscales and indeed, the PHQ-2, have a significantly higher correlation with the WSAS than the GAD-7. In sum the results suggest that both the PHQ-9 and the GAD-7 can be used reliably when considered as two separate scales, but there may also be some use in recognizing the possible cognitive/affective and somatic subscales of each. The subscale information may prove useful for clinical purposes. For example, Elhai, *et al.* [17] found that, the somatic items of a depression measure were significantly more related than the cognitive-affective items to Post Traumatic Stress Disorder (PTSD) factors in Canadian military veterans.

References

1. Löwe B, Spitzer RL, Williams JB, Mussell M, Schellberg D, et al. (2008) Depression, anxiety and somatization in primary care: syndrome overlap and functional impairment. *Gen Hosp Psychiatry* 30: 191-199. Crossref]

2. Olfson M, Shea S, Feder A, Fuentes M, Nomura Y, et al. (2000) Prevalence of anxiety, depression, and substance use disorders in an urban general medicine practice. *Arch Fam Med* 9: 876-883. [Crossref]

3. Spitzer RL, Kroenke K, Williams JBW, Lowe B (2006) A brief measure for assessing generalised anxiety disorder: The GAD-7. *Arch Intern Med* 166: 1092-1097.

4. Jordan P, Shedden-Mora MC, Löwe B (2017) Psychometric analysis of the Generalized Anxiety Disorder scale (GAD-7) in primary care using modern item response theory. *PLoS One* 12: 0182162. [Crossref]

5. Kroenke K, Spitzer RL, Williams JBW (2001) The PHQ-9: Validity of a brief depression severity measure. *J Gen Intern Med* 16: 606-613.

6. Kroenke K, Spitzer RL, Williams JB, Löwe B (2009) An ultra-brief screening scale for anxiety and depression: the PHQ-4. *Psychosomatics* 50: 613-621. [Crossref]

7. Ryan TA, Bailey A, Fearon P, King J (2013) Factorial invariance of the Patient Health Questionnaire and Generalised Anxiety Disorder Questionnaire. *Br J Clin Psychol* 52: 438-449.

8. Beard C, Hsu KJ, Rifkin LS, Busch AB, Björgvinsson T (2016) Validation of the PHQ-9 in a psychiatric sample. *J Affect Disord* 193: 267-273. [Crossref]

9. Elhai JD, Contractor AA, Tamburrino M, Fine TH, Prescott MR, et al. (2012) The factor structure of major depression symptoms: A test of four competing models using the Patient Health Questionnaire-9. *Psychiatry Research* 199: 169-173.

10. Kertz S, Bigda-Peyton J, Bjorgvinsson T (2013) Validity of the Generalised Anxiety Disorder-7 scale in an acute Psychiatric population. *Clin Psychol Psychother* 20: 456-464.

11. Beard C, Bjorgvinsson T (2014) Beyond generalised anxiety disorder: Psychometric properties of the GAD-7 in a heterogeneous psychiatric sample. *J Anxiety Disord* 28: 547-552.

12. Toft T, Fink P, Oernboel E, Kristensen K, Frostholm L, et al. (2005) Mental disorders in primary care; prevalence and co-morbidity among disorders. *Psychol Med* 67: 596-601.

13. Mokken RJ (1971) A theory and procedure of scale analysis. De Gruyter, New York.

14. Marks I (1986) The Work and Social Adjustment Scale. Institute of Psychiatry, London.

15. Rosseel Y (2012) lavvan: An R package for structural equation modelling. *Journal of Statistical Software* 48: 1-36.

16. Van der Ark LA (2007) Mokken scale analysis in R. *Journal of Statistical Software* 20: 1-19.

17. Elhai JD, Contractor AA, Palmieri PA, Forbes D, Richardson JD (2011) Exploring the relationship between underlying dimensions of posttraumatic stress disorder and depression in a national, trauma-exposed military sample. *J Affect Disord* 133: 477-480. [Crossref]