Psychiatry Research 271 (2019) 532–534

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Short communication

PHQ-9: One factor or two?

Luke Boothroyd[a], David Dagnan[b], Steven Muncer[c,d,*]

[a] Clinical Psychology, TEWV NHS Trust, Middlesbrough, United Kingdom
[b] Clinical Psychology, Cumbria NHS Trust, Carlisle, United Kingdom
[c] Department of Psychology, University of Durham, Durham, United Kingdom
[d] Clinical Psychology, Teesside University, Middlesbrough, United Kingdom

ABSTRACT

The PHQ-9 has been found to be reliable and valid, but with both a single and two-factor structure suggested in the literature. This measure has not yet been subject to psychometric investigation using Mokken scale analysis. Confirmatory Factor Analysis (CFA) and Mokken analysis were performed on retrospective data from an NHS Trust. CFA found a two-factor structure was a significantly better fit; however, the factors were highly correlated. Mokken analysis suggested that a single scale was viable. The findings support recent research by Gonzalez-Blanch et al. (2018) in suggesting a one factor model is most appropriate for the PHQ-9

1. Introduction

The PHQ-9 is a nine item self-rating measure of depression severity developed by Spitzer et al. (2001). The nine items on the PHQ-9 correspond to the nine DSM-IV diagnostic criteria for depression. The validity of the PHQ-9 has been investigated in two large scale primary care studies across the United States (Spitzer et al., 2001) The PHQ-9 displayed good internal reliability in both studies ($\alpha = 0.89$ and $\alpha = 0.86$ respectively). A single factor model was also found by Ryan et al. (2013) in PHQ-9 data ($N = 23{,}672$) gathered from a London Improving Access to Psychological Therapies (IAPT) service; confirmatory factor analysis (CFA) found a single factor model to be a good fit for the completion of the PHQ 9 face-to-face (SRMR = 0.04, RMSEA = 0.09, CFI = 0.94).

The PHQ-9, however, has not always been found to fit a single factor model. For example, Titov et al. (2011) found a single factor had a poor fit in a sample of 172 depressed patients. Furthermore, Richardson and Richards (2008) found a two factor structure when using exploratory factor analysis in a study of 2,570 spinal injury patients. Beard et al. (2016) studied 1,023 psychiatric participants who completed the PHQ-9 at admission and discharge from an outpatient programme. 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. (2012) 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 = 0.96, TLI = 0.94, RMSEA = 0.05) fitted the data better than a single factor model ($X^2 = 317.71$, $p < .001$, CFI = 0.94, TLI = 0.91, RMSEA = 0.06). 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 factor. More recently still Gonzalez-Blanch et al. (2018) found that both one and two factor models provided reasonable fit. They argued, however, that as the two factors were highly correlated, for the sake of parsimony, a one factor model should be preferred.

Although the PHQ-9 has been found to be a psychometrically valid and reliable tool, it is unclear from CFA whether a single or two-factor structure is best. 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 (Mokken, 1971). 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. In the present study we use both Mokken scaling and CFA to investigate the structure of the PHQ-9. The use of both classical psychometric methods and item response theory should provide a fuller picture of the overall structure

* Corresponding author at: Clinical Psychology, Teesside University, Middlesbrough, United Kingdom.
E-mail address: s.muncer@tees.ac.uk (S. Muncer).

https://doi.org/10.1016/j.psychres.2018.12.048
Received 27 April 2018; Received in revised form 6 December 2018; Accepted 6 December 2018
Available online 07 December 2018
0165-1781/

2. Method

2.1. Sample

The sample consisted of 4,348 adult males (36%) and 7,603 adult females (64%) who had completed the PHQ-9. The mean age for the full sample was 43.23 years (SD = 15.48; range 17–93); male mean age = 43.28 (SD = 15.17; range 17–93), female mean age = 43.14 (SD = 15.64; range 17 – 92). Data was retrieved from a database for all service users accessing a primary care service for people with depression and anxiety in the north of England between February 2009 and August 2015. The first contact with this service by telephone, at which point the PHQ-9 was collected. It has been established previously that data collected by telephone is acceptable (Ryan et al., 2013).

2.2. Data analysis

Ordinal alpha was used to calculate the reliability of the measures, which modifies Cronbach's alpha to take into account the ordinal nature of the data. Confirmatory factor analysis (CFA) was carried out using the SEM Package in R Commander. Mokken analysis was used to further understand the latent traits of the scales using the Mokken package in R (van der Ark, 2012). 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.

3. Results

The item means and totals for the scale can be seen in Table 1. The ordinal alpha for the PHQ-9 scale was 0.9. For CFA a Comparative Fit Index (CFI) of more than 0.95 and a Root Mean Square Error of Approximation (RMSEA) of less than 0.06 have been taken as indicating a good fit. The one factor model was found to be a moderate fit (Satorra-Bentler $\chi^2 = 2586.53$, $df = 27$, CFI = 0.92, RMSEA = 0.10). The two-factor model, based upon a cognitive-affective factor and a somatic factor, was found to be a significantly better fit (Satorra-Bentler $\chi^2 = 1692.52$, $df = 26$, CFI = 0.95, RMSEA = 0.08) than a single factor model ($\Delta X^2 = 894.01$, $p < .01$). There was a high correlation between the two factors ($r = 0.87$), as both Gonzalez-Blanch et al. (2018) and Elhai (2012) also found. The two scales were significantly correlated ($r = 0.68$, $p < .001$). The ordinal alpha for the cognitive scale was 0.86 and 0.83 for the somatic scale.

Mokken scale analysis examined the unidimensionality of the items on the PHQ-9. The automated item selection procedure of the Mokken package (van der Ark, 2012), which selects items that meet Mokken criterion, was used. The H_i values of all nine items for the full data set were above the recommended threshold for retaining items ($H_i > 0.3$) and were deemed to be sufficiently homogenous and unidimensional based on their H_i values and Loevinger's H for the scale of 0.47 ($SE = 0.004$). Each item within the full data sample was sufficiently homogeneous to demonstrate the PHQ-9 is measuring the same underlying construct with a hierarchy of responses on a single scale. A strong scale is evident with an H value of 0.5 and above, moderate scales $H = 0.4 – 0.5$ and weak scales $H = 0.3 – 0.4$ (Molenaar et al., 2000). In this case the PHQ-9 is a moderate scale. It should also be noted that there were no violations of monotonicity and the invariant item ordering statistic H^T was 0.31 suggesting a low but acceptable level of invariant order. Furthermore, nine items were automatically selected for a scale if the default criterion was raised to 0.4. (The interested reader is referred to Stochl et al. (2012) for more details on these concepts)

Mokken analysis can be used in a confirmatory way to check whether a proposed scale is acceptable. Both of the scales identified by CFA would be considered moderate to strong from a Mokken standpoint; with H values of 0.59 for the cognitive scale and 0.46 for the somatic scale and with acceptable ordinal alphas of 0.86 and 0.83.

4. Discussion

The PHQ-9, overall, was found to have excellent reliability for both the one and two scale version. The two-factor model (somatic and cognitive-affective) has a significantly better fit than a single factor model when examined with CFA, as others have also found. The Mokken scale analysis revealed the PHQ-9 to be a moderately strong scale which retained all nine items, but the separate scales were also strong. The unidimensional scale results and the correlation of the two factors from the CFA suggest that the PHQ-9 total score is an acceptable representation of its results.

Although there is some evidence from a psychometric standpoint that both the somatic factor and cognitive affective factor can be identified, one of the weaknesses of the present study is that it offers no proof for the validity or importance of their use. Future research which examines the separate factors in more detail might provide clinicians and services with data to support their use. However it is worth noting that in this study and also those of Gonzalez-Blanch et al. (2012) and Elhai et al (2011) the correlation between the two factors was over 0.85. Indeed, even in samples where one might expect a greater separations of factors such as Richardson and Richards (2008) with spinal injury patients the weighted correlation was still over 0.7, The evidence suggests that separately assessing factors will not provide any useful information for the majority of patients.

Table 1
Mokken Scale analysis of Full PHQ-9 and somatic and cognitive factors.

Item	Mean (SD)	*H* of full PHQ-9	*H* of somatic factor	*H* of cognitive factor
Little interest	1.75 (1.03)	0.72		0.72
Feeling depressed	1.94 (0.97)	0.72		0.72
Trouble sleeping	2.07 (1.06)	0.50	0.47	
Feeling tired	2.07 (1.00)	0.52	0.48	
Poor appetite	1.61 (1.17)	0.49	0.44	
Feeling bad about oneself	1.86 (1.09)	0.59		0.59
Trouble concentrating	1.57 (1.11)	0.54	0.47	
Slow or restless	1.10 (1.11)	0.47	0.43	
Suicidal ideation	0.57 (0.89)	0.62		0.62
Total scale	14.53 (6.50)	0.47		

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2018.12.048.

References

Beard, C., Hsu, K.J., Rifkin, L.S., Busch, A.B., Bjorgvinsson, T., 2016. Validation of the PHQ-9 in a psychiatric sample. J. Affect. Disord. 193, 267–273. https://doi.org/10.1016/j.jad.2015.12.075.

Elhai, J.D., Contractor, A.A., Tamburrino, M., Fine, T.H., Prescott, M.R., Shirley, et al., 2012. The factor structure of major depression symptoms: a test of four competing models using the Patient Health Questionnaire-9. Psychiat. Res. 199, 169–173. https://doi.org/10.1016/j.psychres.2012.05.018.

Gonzalez-Blanch, C, Medrano, L.A., Munoz-Navarro, R., Ruiz-Rodriguez, P., Moriana, J.A., Limonero, J.T., et al., 2018. Factor structure and measurement invariance across various demographic groups and over time for the PHQ-9 in primary care patients in Spain. PloS ONE 13 (2), e0193356.

Mokken, R.J., 1971. A Theory and Procedure of Scale Analysis. De Gruyter, Berlin.

Molenaar, I.W., Sijtsma, K., Boer, P., 2000. MSP5 for windows: a program for Mokken scale analysis for polytomous items. ProGAMMA., Gronnigen.

Richardson, E.J., Richards, J.S., 2008. Factor structure of the PHQ-9 screen for depression across time since injury among persons with spinal chord injury. Rehabil. Psychol. 53, 243–249.

Ryan, T.A., Bailey, A., Fearon, P., King, J., 2013. Factorial invariance of the patient health questionnaire and generalised anxiety disorder questionnaire. Brit. J. Clin. Psychol. 52 (4), 438–449. https://doi.org/10.1111/bjc.12028.

Spitzer, R.L., Williams, J.B.W., Kroenke, K., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Internl. Med. 16, 606–613. https://doi.org/10.1046/j.1525-497.2001.016009606.x.

Stochl, J., Jones, P.B., Croudace, T.J., 2012. Mokken scale analysis of mental health and well-being item responses: a non-parametric IRT method in empirical research for applied health researchers. BMC Med. Res. Methodol. 12, 74–90.

Titov, N., Dear, B., McMillan, D., Anderson, T., Zou, J., Sunderland, M., 2011. Psychometric comparison of the PHQ 9 and BDI II for measuring response during treatment of depression. Cogn. Behav. Therapy 40, 126–136.

van der Ark, A., 2012. New developments in Mokken scale analysis. J. Stat. Softw. 48 (5), 1–27.