Journal of Affective Disorders 193 (2016) 267–273

Contents lists available at ScienceDirect

Journal of Affective Disorders

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

Research paper

Validation of the PHQ-9 in a psychiatric sample

C. Beard *, K.J. Hsu, L.S. Rifkin, A.B. Busch, T. Björgvinsson

McLean Hospital/Harvard Medical School, United States

ARTICLE INFO

Article history:
Received 30 September 2015
Received in revised form
7 December 2015
Accepted 30 December 2015
Available online 31 December 2015

Keywords:
Assessment
Depression
PHQ-9
Validation
Psychometric

ABSTRACT

Background: The PHQ-9 was originally developed as a screener for depression in primary care and is commonly used in medical settings. However, surprisingly little is known about its psychometric properties and utility as a severity measure in psychiatric populations. We examined the full range of psychometric properties of the PHQ-9 in patients with a range of psychiatric disorders (i.e., mood, anxiety, personality, psychotic).

Methods: Patients ($n=1023$) completed the PHQ-9 upon admission and discharge from a partial hospital, as well as other self-report measures of depression, anxiety, well-being, and a structured diagnostic interview.

Results: Internal consistency was good ($\alpha=.87$). The PHQ-9 demonstrated a strong correlation with a well-established measure of depression, moderate correlations with related constructs, a weak correlation with a theoretically unrelated construct (i.e., disgust sensitivity), and good sensitivity to change, with a large pre- to post-treatment effect size. Using a cut-off of ≥ 13, the PHQ-9 demonstrated good sensitivity (.83) and specificity (.72). A split-half exploratory factor analysis/confirmatory factor analysis suggested a two-factor solution with one factor capturing cognitive and affective symptoms and a second factor reflecting somatic symptoms. Psychometric properties did not differ between male and female participants.

Limitations: No clinician-rated measure of improvement, and the sample lacked ethnoracial diversity.

Conclusions: This first comprehensive validation of the PHQ-9 in a large, psychiatric sample supported its use as a severity measure and as a measure of treatment outcome. It also performed well as a screener for a current depressive episode using a higher cut-off than previously recommended for primary care samples.

1. . Introduction

The Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001) was developed as a screener for depression in primary care. It is an appealing tool due to its brevity (9 items), ease of completion for the patient (e.g., same response options for each item), ease of scoring and interpretation, and public availability. Additionally, the PHQ-9 assesses each of the diagnostic criteria for a Major Depressive Episode (MDE). Scores range from 0 to 27, and a cut-score of ≥ 10 has been recommended for detecting cases of current MDE (Kroenke and Spitzer, 2002). Over 100 studies have examined the PHQ-9 for use in primary care and beyond (for a review, see Kroenke et al. (2010)). For example, the PHQ-9 has been validated in numerous specific medical populations, as well as in general population samples (Gelaye et al., 2014; Kiely and Butterworth, 2015; Kocalevent et al., 2013; Martin et al., 2006). Studies have spanned multiple countries and languages, as well as multiple formats (e.g., automated telephone-based administration, touch-screen computer).

In stark contrast, few studies have validated the PHQ-9 in psychiatric samples. Patients treated in mental health settings have unique characteristics, including more severe and comorbid symptoms compared to general medical or population samples. Therefore, it is crucial to examine the PHQ-9's psychometric properties and utility as a severity measure, in addition to a screener, specifically in psychiatric samples. Although no studies to date have comprehensively evaluated the PHQ-9, four studies have assessed at least one psychometric property in a psychiatric sample (Inoue et al., 2012; Pilkonis et al., 2014; Ryan et al., 2013; Titov et al., 2011). In these studies, estimates of internal consistency have been acceptable to good ($\alpha=.74$: Titov et al., 2011; $\alpha=.81$: Pilkonis et al., 2014), and the PHQ-9 has shown moderate to strong associations with related measures of depression and anxiety (Inoue et al., 2012; Pilkonis et al., 2014; Ryan et al., 2013). Studies examining the PHQ-9's sensitivity to change found it was

* Correspondence to: McLean Hospital Behavioral Health Partial Hospital, 115 Mill St Mail Stop 113 Belmont, MA 02478, United States.
E-mail address: cbeard@mclean.harvard.edu (C. Beard).

http://dx.doi.org/10.1016/j.jad.2015.12.075
0165-0327/

similar to the Beck Depression Inventory-II (Titov et al., 2011), but it defined fewer patients as recovered compared to the Center for Epidemiological Studies of Depression Scale (CESD) and Patient Reported Outcomes Measurement Information System (PROMIS; Pilkonis et al., 2014). The only study examining screening properties compared the PHQ-9 to a psychiatrist's diagnosis in an outpatient clinic in Japan. The authors suggested a cut-off score of 13/14 and concluded that the PHQ-9 was appropriate to use for screening and assessing depressive symptom severity (Inoue et al., 2012). Finally, studies testing the proposed factor structure of the PHQ-9 had inconsistent results, with one supporting the proposed unidimensional factor structure (Ryan et al., 2013), and one failing to confirm it (Titov et al., 2011).

Although these few studies suggest the PHQ-9 may be an adequate measure of depression severity in psychiatric samples, its utility as a screening instrument and sensitive treatment outcome measure is unclear. In particular, the one study that examined the PHQ-9's sensitivity and specificity relied on a single psychiatrist's diagnosis as a gold standard rather than a structured diagnostic interview (Inoue et al., 2012). Additionally, several studies utilized samples that are not representative of real-world psychiatric settings (e.g., excluded individuals with a history of bipolar or psychotic disorders (Pilkonis et al., 2014)), excluded individuals with severe depression or suicidal ideation (Titov et al., 2011). Such excluded patients are the norm in psychiatric settings, and thus a more representative sample is required to validate the PHQ-9. Finally, the two studies that did include patients from standard care settings were conducted in Japan (Inoue et al., 2012) and the United Kingdom (Ryan et al., 2013) and had very narrow aims. Inoue et al. (2012) only examined the PHQ-9's screening properties, and Ryan et al. (2013) focused on factorial invariance of different formats of the scale.

In sum, despite the wealth of data supporting the PHQ-9 and its widespread use in medical settings, no study has adequately validated its use in a psychiatric sample. This gap in the literature is particularly surprising given that the PHQ-9 is the recommended measure of depression severity in the Diagnostic Statistical Manual-5 (American Psychiatric Association, n.d.). The current study sought to provide the first comprehensive evaluation of the PHQ-9 in a large psychiatric sample. We utilized data from a partial hospital treatment program that administered the PHQ-9 as part of routine clinical care. We examined internal consistency, convergent validity with other measures of depression and related constructs, discriminant validity with constructs not theoretically associated with depression, sensitivity to change following partial hospital treatment, and sensitivity and specificity in detecting cases of a MDE in comparison to a structured diagnostic interview. We compared the PHQ-9's performance on these indices to a well-validated depression measure, the CESD-10 (Andresen et al., 1994). Based on the wealth of studies validating the PHQ-9 in other settings, we expected to obtain good reliability and convergent validity. We expected a higher cut-score would be required to achieve adequate screening properties given the severe symptom levels and comorbidity present in real world psychiatric settings.

2. Method

2.1. Participants and treatment setting

Participants were patients receiving treatment at the Behavioral Health Partial Hospital at McLean Hospital from July of 2013 to March of 2015 ($n=1211$). The only patients excluded were those who did not consent for their clinical data for be used for research purposes ($n=188$; final sample $n=1023$). As described elsewhere (Beard and Björgvinsson, 2014), the partial hospital program delivers cognitive behavioral therapy (CBT), pharmacotherapy, and aftercare planning to patients suffering from a wide range of psychiatric disorders, principally mood, anxiety, personality, and psychotic disorders. Approximately half of patients are referred from an inpatient hospital, while the other half are referred from outpatient care or the community. Patients attend five 50-min groups each day, five days per week (Monday–Friday).

Table 1
Demographic and clinical characteristics ($n=1023$).

Demographic characteristics	*N*	(%)
Female	534	(52.2%)
Age (*M*, SD)	34.30	(13.36)
Ethnicity		
Non-Latino/a	992	(96.97%)
Latino/a	31	(3.03%)
Race		
White	873	(88.00%)
Black/African American	24	(2.42%)
Asian	38	(3.83%)
American Indian/Alaskan Native	2	(.00%)
Native Hawaiian/Pacific Islander	1	(.00%)
Multi-racial	34	(3.43%)
Other/Unknown	31	(3.13%)
Marital Status		
Single	628	(61.63%)
Married/Living with Partner	261	(25.61%)
Divorced/Widowed	130	(12.76%)
Highest Level of Education		
High School/GED	91	(8.91%)
Some college	377	(36.92%)
Post-undergraduate	553	(54.16%)
Clinical Characteristics[a]	N	(% out of 850)
Current Major Depressive Episode	514	(60%)
Current Major Depressive Disorder	411	(48%)
Bipolar Disorder I	168	(20%)
Bipolar Disorder II	33	(4%)
Social Anxiety Disorder	251	(30%)
Generalized Anxiety Disorder	231	(27%)
Panic Disorder	199	(23%)
Post-Traumatic Stress Disorder	105	(12%)
Obsessive Compulsive Disorder	91	(11%)
Psychotic Disorder	50	(6%)

Note: Diagnostic percentages exceed 100% due to comorbidity.

[a] 173 patients did not complete a MINI diagnostic interview for various reasons (e.g., too acute, transferred to inpatient, scheduling difficulties). We include these patients in the total sample analyses that do not involve diagnostic groups.

Participants were primarily single, White, and middle-age (see Table 1). The average duration of treatment in the sample for this study was 10.7 (SD=4.63) days. The most common current DSM-IV Axis I diagnosis from a structured interview (see Section 2.2) was Major Depressive Disorder, followed by Social Anxiety Disorder (SAD).

2.2. Measures

Mini International Neuropsychiatric Interview (MINI; Sheehan et al., 1998) is a structured interview assessing for DSM-IV Axis I disorders. The MINI has strong reliability and validity in relation to the Structured Clinical Interview for DSM-IV, with inter-rater reliabilities ranging from kappas of .89–1.0 (Sheehan et al., 1998). The MINI was administered by doctoral practicum students and interns in clinical psychology who received weekly supervision by a postdoctoral psychology fellow. Training included reviewing administration manuals and completing mock interviews. All clinicians were required to pass a final training interview with their supervisor before administering MINIs for the program. MINI raters meet bi-annually to rate an audio recording of a MINI

interview. Reliability ratings yielded near perfect agreement (Cohen's Kappa=.911) on diagnoses.

Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001) is a 9-item self-report measure which is used to assess depression severity and criteria for a major depressive episode (MDE). Items assess for symptoms of depression (e.g., "*little interest or pleasure in doing things*") and response anchors range temporally from 0 *(not at all)* to 3 *(nearly every day)*.

Center for the Epidemiological Studies of Depression-10 (CESD-10; Andresen et al., 1994) is a widely-used instrument measuring symptoms of depression (e.g., "*I felt depressed*"), with response anchors ranging from 0=*rarely or none of the time (less than 1 day)* to 3=*most or all of the time (5–7 days)*. The CESD-10 had good internal consistency (α=.85).

The 7-item Generalized Anxiety Disorder Scale (GAD-7; Spitzer et al., 2006) is a self-report questionnaire that assesses symptoms of general anxiety (e.g., "*trouble relaxing*") according to a 4-point Likert type scale, from 0 (*not at all*), to 3 (*nearly every day*). The GAD-7 has demonstrated good reliability and construct validity (Kertz et al., 2013; Löwe et al., 2008; Spitzer et al., 2006), including in this partial hospital setting (Beard and Bjorgvinsson, 2014; Kertz et al., 2013). The GAD-7 had good internal consistency (α=.88).

Schwartz Outcome Scale (SOS; Blais et al., 1999) is a well-validated single-factor, 10-item measure designed to examine the broad domain of psychological health (Young et al., 2003). Each item assesses for psychological well-being (e.g., "*My life is according to my expectations*") using a 7-point Likert scale from 0 (*Never*) to 6 (*All or nearly all of the time*). Internal consistency of the SOS was excellent (α=.93).

The Disgust Propensity and Sensitivity Scale – Revised (DPSS-R; Fergus and Valentiner, 2009) is a 12-item self-report questionnaire measuring disgust propensity, the frequency of disgust experiences (e.g. "*I avoid disgusting things*"), and disgust sensitivity, the emotional impact of disgust experiences (e.g. "*When I feel disgusted, I worry that I might pass out*"). Patients were asked to rate each item on how often it is true to them on a scale ranging from 1 ("*never*") to 5 ("*always*"). Higher scores on this measure reflect higher levels of disgust propensity and sensitivity. The DPSS demonstrated good internal consistency (α=.87).

Clinical Global Impression Scale-Improvement (CGIS; Guy, 1976) is a 7-point scale assessing a patient's improvement during treatment (or lack thereof) compared to baseline status. In the current study, patients rated their own improvement at discharge from the partial hospital using the scale from "*very much improved*" to "*very much worse*". Patient ratings have been found to have moderate correlations with provider ratings (ICC=.65) and comparable validity (Forkmann et al., 2011).

2.3. Procedure

Upon admission to the program, patients were informed that they would complete computerized questionnaires to assess their symptoms. Only patients who provided informed written consent for their responses to be included in research are included in the current report. The local Institutional Review Board approved all study procedures. The pre-treatment and post-treatment assessments included the PHQ-9, CESD-10, GAD-7, SOS, and DSPS-R. Patients typically completed the MINI on their second day in the program and the CGIS at discharge. Study data were collected and managed using REDCap electronic data capture tools hosted at McLean Hospital. REDCap (Research Electronic Data Capture) is a secure, web-based application designed to support data capture for research studies (Harris et al., 2009).

2.4. Analyses

SPSS version 21.0 (Spss, 2012) and MPlus version 6.12 (Muthén and Muthén, 2011) were used for all statistical analyses. We conducted the following analyses for the total sample and separately by gender. Internal consistency was estimated using Cronbach's alpha. We assessed convergent validity via correlations with a well-established measure of depression (CESD-10) and with measures of closely related constructs (anxiety (GAD-7) and well-being (SOS)). To further examine convergent validity, we compared PHQ-9 scores for patients with and without a current MDE (diagnosed by the MINI) using an independent samples *t*-test. To examine discriminant validity, we examined the correlation between the PHQ-9 and a theoretically unrelated construct of disgust propensity (DSPS-R). To examine sensitivity to change, we calculated effect sizes (Cohen's *d*) for pre- to post-treatment changes and conducted paired-samples *t*-tests. We also examined the number of patients achieving a reliable change on the PHQ-9 as defined by a Reliable Change Index $\geq$1.96 (Jacobson and Truax, 1991). Using .78 as the test–retest reliability (Delgadillo et al., 2011), patients needed to demonstrate a reduction of at least 8.6 points to achieve a reliable change. Additionally, we compared pre–post changes on the PHQ-9 between patients who self-reported global improvement following treatment (i.e., 'much' or 'very much' improved at discharge on the Clinical Global Improvement Scale-self-report) and those who did not improve. Sensitivity, specificity, and positive and negative predictive values were calculated based on the initial recommended cut-off of $\geq$10 and a higher cut-off of $\geq$13 based on one prior depressed sample (Inoue et al., 2012). We also examined the properties using the diagnostic algorithm, which follows the Diagnostic and Statistical Manual of Mental Disorders, fourth edition text revision (DSM-IV-TR; American Psychiatric Association, 2000) and fifth edition (DSM-5; American Psychiatric Association, 2013). Specifically, patients must endorse a response of 2 ("*more days than not*") on either the depressed mood item or the anhedonia item, and at least five symptoms total with a response of 2 or greater.

To identify and validate the factor structure underlying the PHQ-9, we conducted a split-half exploratory factor analysis (EFA)/confirmatory factor analysis (CFA), given that this is the first study to examine the PHQ-9 in a large psychiatric sample. We first conducted an EFA using maximum likelihood estimation with PHQ-9 items at baseline to examine the structure of the scale in half of our participants. Oblique rotations (geomin) were performed due to the correlated nature of the factors. We selected factors based on eigen values $\geq$1, the scree plot, and the factor pattern loadings. Based on results of this EFA, we then conducted a CFA on the other half of participants using the suggested factor structure. We examined standard CFA goodness-of-fit indices and recommended cut-offs (Quintana and Maxwell, 1999) including chi-square, comparative fit index (CFI > .90) and Root Mean Square Error of Approximation (RMSEA < .10). Modification indices were also considered where appropriate based on existing convention (Kline, 2005; MacCallum, 1995): i.e., (1) any modification of a model must be theoretically justifiable; (2) modifications must be few in number; and (3) modifications should be minor.

To examine differences in the factor structure underlying the PHQ-9 between gender, we conducted a multigroup confirmatory factor analysis. This systematic procedure allows for establishment of the validity of a measure across groups (i.e., measurement invariance). Testing for measurement invariance is a multi-tiered process that utilizes constraints from previous levels for each subsequent model, in a stepwise fashion. Measurement invariance may either be full or partial; full invariance suggests that all constraints at a specific level of testing are invariant between groups while partial invariance allows for some relaxation of constraints

to improve model fit. We first fit the model identified above separately in each group, before fitting the model in both groups, allowing all parameters to be free; this model, with freed parameters, serves as a baseline model for comparison and is a test of configural invariance. By constraining factor loadings across groups for the suggested model, metric invariance can be assessed. Metric invariance implies that the measure and underlying construct(s) are reflected similarly by the items across groups. Thus, metric non-invariance suggests that the measure examines different constructs across groups. Scalar invariance suggests that the regression intercepts of items onto their respective factors are similar; scalar non-invariance may reflect group bias to answer higher or lower on an item than expected based on other structural parameters. Scalar invariance is tested through constraining item intercepts between groups. Metric invariance and scalar invariance suggest the measure shows measurement invariance; that is, on the item level the measure performs similarly across groups. Further testing of the measure looks at a more structural level, examining the latent factors associated with the measure, with measurement non-invariance at these steps possibly suggesting differences on a theoretical level between groups. In these instances, measurement non-invariance may indicate that constructs differ across groups and that these differences are being reflected by the measure between the two groups. Invariance of factor variances signifies that factors have equivalent variances across groups and is tested by constraining latent factor variances across groups. Invariance of factor covariances indicates that the associations between latent variables are the same across groups and is tested by constraining latent factor covariances to be equal across groups. Constraining latent factor means across groups tests invariance of factor means; this invariance suggests that mean factor scores are similar across groups. Accordingly, testing for measurement invariance examines a variety of possible differences across gender on the PHQ-9 in a systematic, stepwise fashion.

Models are compared through the CFI and RMSEA fit indices suggested by Chen (2007). Changes of −.010 or greater for CFI and .015 or greater for RMSEA were utilized as cutoffs points; differences between the previous model and the current model exceeding the cutoffs suggest that the assumptions of the current model (e.g., metric invariance) do not hold true across groups. If full invariance is not found, modification indices can be used to determine which parameters may be relaxed to test for partial invariance. Partial invariance still supports some degree of invariance in the measure across groups within that model level (e.g., metric or scalar invariance) and allows for progression in testing of measurement invariance (Chen, 2007).

3. Results

3.1. Means and internal consistency

We first examined demographic differences on the PHQ-9 at baseline. Women scored 2.1 points higher than men, $F(1,1021)=26.4$, $p<.001$. No other demographic differences were observed. The sample mean was 14.5 (SD=6.56) at pre-treatment and 9.89 (SD=5.95) at post-treatment. Internal consistency at pre- and post-treatment was good (Total sample: $\alpha=.87$ at both time points; Females: T1 $\alpha=.85$; T2 $\alpha=.86$; Males: T1 $\alpha=.88$; T2 $\alpha=.89$).

3.2. Construct validity

As expected, higher scores on the PHQ-9 were associated with higher scores on measures of depression (CESD-10, $r(1010)=.80$, $p<.001$) and anxiety (GAD-7, $r(1015)=.61$, $p<.001$), as well as lower scores for psychological well-being (SOS-10, $r(989)=-.65$, $p<.001$). Additionally, patients who met criteria for a current MDE scored significantly higher on the PHQ-9 than patients without a current MDE, $t(847)=21.93$, $p<.001$. As expected, the PHQ-9 had a relatively weak association with the measure of discriminant validity, disgust propensity (DPSS-R, $r(304)=.20$, $p<.001$). The pattern and magnitude of correlations and MDE group differences was similar across genders.

3.3. Sensitivity to change

We examined the PHQ-9's sensitivity to change over the course of partial hospital treatment in patients who reported clinical symptoms of depression at admission (> 10) and who completed discharge assessments ($n=576$). Patients reported significant pre–post changes on the PHQ-9 with a large effect size ($t(575)=24.97$, $p<.001$, $d=1.13$ (95% CI:.77–1.59)). This was comparable to the effect size from the CESD-10 ($t(575)=26.38$, $p<.001$, $d=1.19$ (95% CI:.79–1.64)). Effect sizes were comparable across genders (Females $d=1.16$; Males $d=1.12$).

Patients who achieved a reliable change on the PHQ-9 ($n=175$, 30%) reported less severe depression and anxiety, as well as increased well-being, at post-treatment compared to patients who did not achieve a reliable change (PHQ-9: $t(574)=13.2$, $p<.001$; GAD-7: $t(569)=7.07$, $p<.001$; SOS: $t(564)=-8.60$, $p<.001$). Patients who self-reported 'much' or 'very much' improvement on the CGIS had significantly larger reductions on the PHQ-9 than patients who did not report improvement (Improved: $M=7.05$, $SD=5.50$; Not improved: $M=2.62$, SD=4.27; $t(558)=9.07$, $p<.001$). All sensitivity to treatment change analyses revealed nearly identical results for women and men.

3.4. Sensitivity and specificity[1]

Table 2 presents the sensitivity, specificity, positive predictive value, and negative predictive value for three different criteria for the PHQ-9: cut-point ≥ 10, cut-point ≥ 13, and the diagnostic algorithm derived from DSM-IV-TR/DSM-V. Table 2 also includes these data for the recommended CESD-10 cut-point for comparison. The recommended cut-point of ≥ 10 for the PHQ-9 resulted in sensitivity of .93 and specificity of .52 compared to a MINI diagnosed MDE. The higher cut-off of ≥ 13 resulted in sensitivity of .83 and specificity of .72. The diagnostic algorithm resulted in sensitivity of .90 and specificity of .57. The ROC curve analysis (see Fig. 1) estimated the area under the curve to be .85 (95% confidence interval=.83–.88). Area under the curve was similar across genders (Females=.87 (95% confidence interval=.84–.91); Males=.83 (95% confidence interval=.79–.87)). The CESD-10 recommended cut-offs performed similarly, and the ROC curve analysis estimated the area under the curve for the CESD-10 to be .84 (95% confidence interval=.82–.87).

3.5. Factor structure

Based on the scree plot and eigen values, a two-factor structure emerged as the appropriate solution from the EFA, accounting for

[1] We chose to compare the PHQ-9 to an interview rated MDE because individuals with a variety of diagnoses (MDD, Bipolar, Depression NOS) who are currently experiencing a MDE would not be expected to look different on the PHQ9. In other words, the PHQ-9 would not be expected to distinguish between these diagnoses in individuals who are currently depressed, and thus specificity should be poorer for detecting MDD. Consistent with this, using the higher cut-off of PHQ-9 ≥ 13 to detect current MDD as assessed by the MINI, sensitivity was still good (.84), but specificity was poor (.54), and the area under the curve was lower .74 (95% CI:.71–.77) compared to detecting an MDE.

Table 2
Sensitivity and specificity for cut-scores

	Sensitivity	Specificity	PPV	NPV
PHQ-9 ≥ 10	.93	.52	.74	.84
PHQ-9 ≥ 13	.83	.72	.82	.74
PHQ-9 Diagnostic Algorithm	.90	.57	.76	.79
CESD-10 ≥ 10	.97	.33	.68	.88
CESD-10 ≥ 16	.82	.70	.81	.72

Note: Sensitivity=number of patients correctly classified as depressed on PHQ-9/total number of depressed patients (according to MINI); Specificity=number of patients correctly classified as NOT depressed on PHQ-9/total number of non-depressed patients (according to MIN); Positive Predictive Value (PPV)=number of patients correctly classified as depressed on PHQ-9/total number of patients classified as depressed on PHQ-9; Negative Predictive Value (NPV)=number of patients correctly classified as NOT depressed on PHQ-9/total number of patients classified as NOT depressed on PHQ-9; PHQ-9=Patient Health Questionnaire-9; CESD-10=Center for Epidemiological Studies Depression Scale-10-item version.

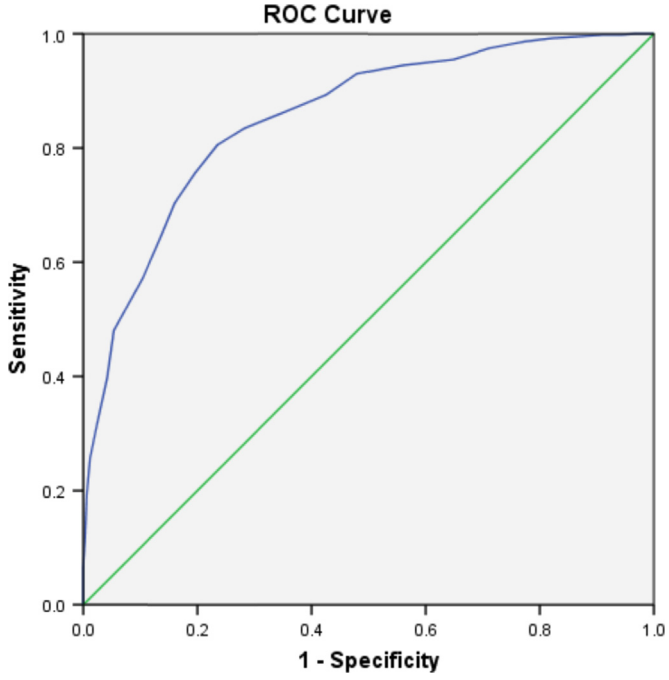

Fig. 1. Receiver operating curve for PHQ-9.

60.20% of the variance. The first factor consisted of cognitive and affective items from the PHQ-9 (e.g., "little interest or pleasure in doing things", "feeling down, depressed or hopeless"), while the second factor consisted of somatic items (e.g., "feeling tired or having little energy", "poor appetite or overeating"). We then applied this model to the other half of the sample in a CFA (see Fig. 2), demonstrating marginal fit to the data, $\chi^2=(26, N=511)=112.08$, $p<.001$, CFI=.96, RMSEA=.08, 90% CI [.07 .10], SRMR=.04. Modification indices for this two factor model suggested significant improvements in model fit would be gained by freeing the covariance between item 7 (concentration difficulty) and item 8 (motor slowing/restlessness). This modification is supported by findings that psychomotor speed is associated with attention (e.g., Lemelin and Baruch, 1998) or may reflect aspects of psychomotor agitation, as argued by others (e.g.,Hepner et al., 2009; Ryan et al., 2013). The inclusion of this single modification index resulted in a model with good fit to the data, $\chi^2=(25, N=511)=74.70$, $p<.001$, CFI=.98, RMSEA=.06, 90% CI [.05 .08], SRMR=.03.

This two-factor model exhibited metric invariance, scalar invariance, invariance of latent variance, invariance of latent covariance, and invariance of latent means (i.e., did not exceed the specific fit index cutoffs for those model progressions), indicating no significant differences between men and women on both a measurement and structural level.

4. Discussion

Although the PHQ-9 was developed as a screener, psychiatric settings may be more likely to use it as a measure of severity and of outcome because patients in these settings are already identified as needing mental health treatment. Indeed, the DSM-5 recommends the PHQ-9 as the severity measure for depression; however, the PHQ-9 has yet to be comprehensively validated in a large psychiatric sample. In the current real-world, psychiatric sample, the PHQ-9 demonstrated strong psychometric properties. Supporting its use as a severity measure, the PHQ-9 demonstrated good convergent validity by its strong correlations with other measures of depression, anxiety, and well-being. It also demonstrated discriminant validity by its weak correlation with a measure of disgust propensity. Regarding sensitivity to treatment, the PHQ-9 revealed a large pre–post-treatment effect size that was comparable to the effect on the CESD-10. Approximately one-third of the sample achieved a reliable change on the PHQ-9, and those who did also reported less anxiety and better well-being at post-treatment compared to those who did not achieve a reliable change. Moreover, those who were classified as improved according to the CGIS showed significantly larger PHQ-9 change scores compared to those who did not improve. Thus, the PHQ-9 appears to be a useful severity measure and is sensitive to changes following brief, partial hospital treatment. Importantly, all of these psychometric properties were similar in women and men.

The PHQ-9 has demonstrated a unidimensional structure in general population samples (e.g., Kocalevent et al., 2013). However, our findings indicate a two-factor structure best fit the data from a usual care psychiatric sample. This two-factor solution distinguishing between cognitive/affective symptoms and somatic symptoms parallels the factor structure of another well-characterized self-report depression symptom questionnaire, the BDI-II (Storch et al., 2004; Whisman et al., 2000). However, these results stand in contrast to findings from a study utilizing samples from randomized clinical trials (RCTs) for depression that excluded severe depression and suicidal ideation (Titov et al., 2011). This discrepancy in findings underscores problems with generalizing findings from RCTs to usual care psychiatric settings and our rationale for conducting the current study. The current findings suggest that for high acuity, heterogeneous populations, the PHQ-9 taps two different symptom dimensions: cognitive–affective symptoms and somatic symptoms. Researchers and clinicians using the PHQ-9 in such settings may find it beneficial to examine responses separately for these two domains (cognitive/affective versus somatic).

Although psychiatric settings may be more likely to use the PHQ-9 as a severity or outcome measure, large clinically heterogeneous settings may still desire a screening instrument in order to reduce the number of diagnoses requiring further assessment. We examined potential clinical cut-off scores and compared screening properties with the CESD-10. The recommended PHQ-9 cut-off for primary care settings of ≥10 yielded adequate sensitivity, but poor specificity and high false positive rates in detecting a current MDE. Given the increased comorbidity, severity, and acuity of psychiatric samples compared to primary care samples, a higher cut-score is likely needed. Indeed, using a cut-off of ≥13 resulted in better specificity, and sensitivity remained good. As a screener, the PHQ-9 performed similarly to the CESD-10, which also required a higher cut-point in acute, psychiatric samples.

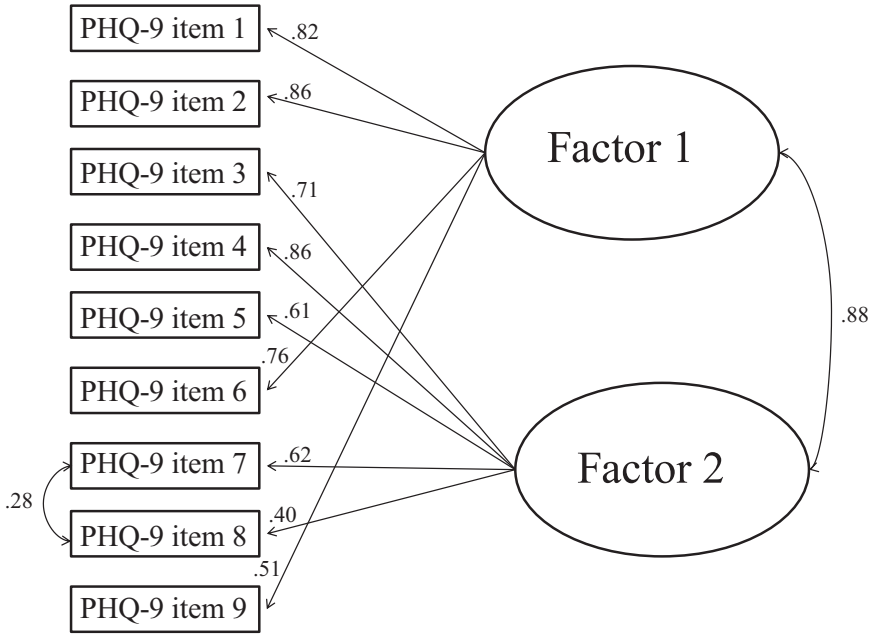

Fig. 2. Confirmatory factor analysis model.

4.1. Limitations

Strengths of the current study include its large, heterogeneous psychiatric sample, evaluation of the full range of psychometric questions related to the use of the PHQ-9 in such settings, and examination of potential gender differences. However, a number of limitations are worth noting. First, our findings regarding the PHQ-9's sensitivity to change are in context of a brief, CBT-based partial hospital. Thus, for some participants, the two-week time period assessed by the PHQ-9 at discharge overlapped to varying degrees with the 2-week time period assessed upon admission, and larger treatment effects might have been obtained with completely non-overlapping time periods. It is encouraging that large effects were observable after only one to two weeks of treatment and with potentially overlapping time periods, as this suggests that the PHQ-9 would be a useful outcome measure for longer-term settings. However, future studies are needed to confirm this. Many providers and settings may desire a shorter time-frame (e.g., past week) in order to measure treatment progress, and future studies validating the PHQ-9 should examine different time frames.

The recommended cut-point will likely generalize to similar acute psychiatric settings (e.g., partial and inpatient hospitals, day programs). However, future studies are needed to determine the appropriate cut-off for other psychiatric settings, such as typical psychiatric outpatient clinics. Second, we did not have a clinician-rated index of improvement. Finally, similar to our overall hospital population, the sample was not diverse in ethno-racial background. Future studies are needed to examine psychometric properties within minority groups.

5. Conclusions

The current study is the first to validate the PHQ-9 in a representative, psychiatric sample. The PHQ-9 performed well as a measure of depression symptom severity and treatment outcome and as a screener using a higher cut-off than has been recommended in primary care settings. Future studies are needed to replicate the two factor structure in other psychiatric samples. Research settings may continue to prefer longer measures of depression severity that assess important symptoms that are not included in the DSMV (see Fried et al. (2016)). Additionally, other measures may be preferable for assessing disaggregated symptoms (e.g., motor restlessness versus slowing). Nonetheless, the PHQ-9 is a useful alternative to more traditional measures of depression in psychiatric settings due to its brevity, assessment of each diagnostic criteria, public availability, and ease of administration, scoring, and interpretation.

References

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders: DSM 5, DSM Library. American Psychiatric Association, Washington, DC.

Andresen, E.M., Malmgren, J.A., Carter, W.B., Patrick, D.L., 1994. Screening for depression in well older adults: evaluation of a short form of the CES-D. Am. J. Prev. Med. 10, 77–84.

Association, A.P., Association, A.P., others, 2000. Diagnostic and statistical manual-text revision (DSM-IV-TRim, 2000). American Psychiatric Association.

Beard, C., Björgvinsson, T., 2014. Beyond generalized anxiety disorder: psychometric properties of the GAD-7 in a heterogeneous psychiatric sample. J. Anxiety Disord. 28, 547–552. http://dx.doi.org/10.1016/j.janxdis.2014.06.002.

Blais, M.A., Lenderking, W.R., Baer, L., deLorell, A., Peets, K., Leahy, L., Burns, C., 1999. Development and initial validation of a brief mental health outcome measure. J. Personal. Assess. 73, 359–373. http://dx.doi.org/10.1207/S15327752JPA7303_5.

Chen, F.F., 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct. Eq. Model.: Multidiscip. J. 14, 464–504. http://dx.doi.org/10.1080/10705510701301834.

Delgadillo, J., Payne, S., Gilbody, S., Godfrey, C., Gore, S., Jessop, D., Dale, V., 2011. How reliable is depression screening in alcohol and drug users? A validation of brief and ultra-brief questionnaires. J. Affect. Disord. 134, 266–271. http://dx.doi.org/10.1016/j.jad.2011.06.017.

Fergus, T.A., Valentiner, D.P., 2009. The disgust propensity and sensitivity scale–revised: an examination of a reduced-item version. J. Anxiety Disord. 23, 703–710. http://dx.doi.org/10.1016/j.janxdis.2009.02.009.

Forkmann, T., Scherer, A., Boecker, M., Pawelzik, M., Jostes, R., Gauggel, S., 2011. The clinical global impression scale and the influence of patient or staff perspective on outcome. BMC Psychiatry 11, 83. http://dx.doi.org/10.1186/1471-244X-11-83.

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are 'good' depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of affective disorders*, *189*, 314-320.

Gelaye, B., Tadesse, M.G., Williams, M.A., Fann, J.R., Vander Stoep, A., Andrew Zhou, X.-H., 2014. Assessing validity of a depression screening instrument in the absence of a gold standard. Ann. Epidemiol. 24, 527–531. http://dx.doi.org/10.1016/j.annepidem.2014.04.009.

Guy, W., 1976. Clinical global impression scale. The ECDEU Assessment Manual for Psychopharmacology – Revised. Volume DHEW Publ. No. ADM 76, 338, pp. 218–222.

Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J.G., 2009. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. 42, 377–381. http://dx.doi.org/10.1016/j.jbi.2008.08.010.
Hepner, K.A., Hunter, S.B., Edelen, M.O., Zhou, A.J., Watkins, K., 2009. A comparison of two depressive symptomatology measures in residential substance abuse treatment clients. J. Subst. Abuse Treat. 37, 318–325.
Inoue, T., Tanaka, T., Nakagawa, S., Nakato, Y., Kameyama, R., Boku, S., Toda, H., Kurita, T., Koyama, T., 2012. Utility and limitations of PHQ-9 in a clinic specializing in psychiatric care. BMC psychiatry 12, 73.
Jacobson, N.S., Truax, P., 1991. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J. Consult. Clin. Psychol. 59, 12.
Kertz, S., Bigda-Peyton, J., Bjorgvinsson, T., 2013. Validity of the generalized anxiety disorder-7 scale in an acute psychiatric sample. Clin. Psychol. Psychother. 20, 456–464. http://dx.doi.org/10.1002/cpp.1802.
Kiely, K.M., Butterworth, P., 2015. Validation of four measures of mental health against depression and generalized anxiety in a community based sample. Psychiatry Res. 225, 291–298. http://dx.doi.org/10.1016/j.psychres.2014.12.023.
Kline, R.B., 2005. Principles and Practice of Structural Equation Modeling, 2nd ed. The Guilford Press, New York.
Kocalevent, R.-D., Hinz, A., Brähler, E., 2013. Standardization of the depression screener patient health questionnaire (PHQ-9) in the general population. Gen. Hosp. Psychiatry 35, 551–555. http://dx.doi.org/10.1016/j.genhosppsych.2013.04.006.
Kroenke, K., Spitzer, R.L., 2002. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr. Ann. 32, 1–7.
Kroenke, K., Spitzer, R.L., Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16, 606–613.
Kroenke, K., Spitzer, R.L., Williams, J.B.W., Löwe, B., 2010. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen. Hosp. Psychiatry 32, 345–359. http://dx.doi.org/10.1016/j.genhosppsych.2010.03.006.
Lemelin, S., Baruch, P., 1998. Clinical psychomotor retardation and attention in depression. J. Psychiatr. Res. 32, 81–88.
Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., Herzberg, P.Y., 2008. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med. Care 46, 266–274. http://dx.doi.org/10.1097/MLR.0b013e318160d093.
MacCallum, R.C., 1995. Model specification: procedures, strategies, and related issues. In: Hoyle, R.H. (Ed.), Structural Equation Modeling: Concepts, Issues, and Applications. Sage, Thousand Oaks, CA.
Martin, A., Rief, W., Klaiberg, A., Braehler, E., 2006. Validity of the brief patient health questionnaire mood scale (PHQ-9) in the general population. Gen. Hosp. Psychiatry 28, 71–77. http://dx.doi.org/10.1016/j.genhosppsych.2005.07.003.
Muthén, L.K., Muthén, B.O., 2011. Mplus User's Guide, Sixth. ed. Muthén & Muthén, Los Angeles, CA.
Pilkonis, P.A., Yu, L., Dodds, N.E., Johnston, K.L., Maihoefer, C.C., Lawrence, S.M., 2014. Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study. J. Psychiatr. Res. 56, 112–119. http://dx.doi.org/10.1016/j.jpsychires.2014.05.010.
Quintana, S.M., Maxwell, S.E., 1999. Implications of recent developments in structural equation modeling for counseling psychology. Couns. Psychol. 27, 485–527. http://dx.doi.org/10.1177/0011000099274002.
Ryan, T.A., Bailey, A., Fearon, P., King, J., 2013. Factorial invariance of the patient health questionnaire and generalized anxiety disorder questionnaire. Br. J. Clin. Psychol. 52, 438–449. http://dx.doi.org/10.1111/bjc.12028.
Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiatry 59, 22–33.
Spitzer, R.L., Kroenke, K., Williams, J.B.W., Löwe, B., 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch. Intern. Med. 166, 1092–1097. http://dx.doi.org/10.1001/archinte.166.10.1092.
Spss, I., 2012. IBM SPSS statistics version 21. Boston, Mass: International Business Machines Corp.
Storch, E.A., Roberti, J.W., Roth, D.A., 2004. Factor Structure, Concurrent Validity, and Internal Consistency of the Beck Depression Inventory, second edition19; , pp. 187–189.
Titov, N., Dear, B.F., 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. Ther. 40, 126–136. http://dx.doi.org/10.1080/16506073.2010.550059.
Whisman, M.A., Perez, J.E., Ramel, W., 2000. Factor structure of the beck depression inventory—second edition (BDI-ii) in a student sample. J. Clin. Psychol. 56, 545–551.
Young, J.L., Waehler, C.A., Laux, J.M., McDaniel, P.S., Hilsenroth, M.J., 2003. Four studies extending the utility of the Schwartz Outcome Scale (SOS-10). J. Pers. Assess. 80, 130–138. http://dx.doi.org/10.1207/S15327752JPA8002_02.