General Hospital Psychiatry 56 (2019) 13–18

Contents lists available at ScienceDirect

General Hospital Psychiatry

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

Research paper

Psychometric properties and clinical utility of brief measures of depression, anxiety, and general distress: The PHQ-2, GAD-2, and K-6

Lauren G. Staples[a,b,*], Blake F. Dear[a,b], Milena Gandy[b], Vincent Fogliati[b], Rhiannon Fogliati[b], Eyal Karin[b], Olav Nielssen[a,c], Nickolai Titov[a,b]

[a] *MindSpot Clinic, Macquarie University, Sydney, Australia*
[b] *eCentreClinic, Department of Psychology, Macquarie University, Sydney, Australia*
[c] *Faculty of Medicine and Health Sciences, Macquarie University, Sydney, Australia*

ARTICLE INFO

Keywords:
Patient Health Questionnaire
Generalized Anxiety Disorder Questionnaire
Kessler Psychological Distress Scale
Internet
Cognitive Behavior Therapy
Treatment change

ABSTRACT

Objective: The nine-item Patient Health Questionnaire (PHQ-9), seven-item Generalized Anxiety Disorder scale (GAD-7), and ten-item Kessler Psychological Distress Scale (K-10) are valid and reliable measures of depression, anxiety and general distress. However, the time required in their administration may limit their use in routine care. This study examines the utility of shorter versions (PHQ-2, GAD-2, and K-6) as screening instruments and measures of treatment response.

Method: Data from research trial participants (n = 993) receiving internet-delivered cognitive behaviour therapy (iCBT) were analysed to establish discriminant validity of the short versions. Mini International Neuropsychiatric Interview (MINI) diagnoses were used as comparators. Criterion group validity, test–retest reliability, internal consistency, and responsiveness to treatment changes were examined. Analyses were replicated using data from patients receiving iCBT in routine care (n = 1389).

Results: Discriminant validity was excellent for the PHQ-2, and acceptable for the GAD-2 and K-6. Acceptable sensitivity and specificity were identified at a threshold of ≥3 for the PHQ-2 and GAD-2, and ≥14 for the K-6. The short versions were sensitive to treatment change.

Conclusion: The PHQ-2, GAD-2 and K-6 are useful screeners and efficient measures of treatment progress and outcomes in routine clinical care.

1. Introduction

Recognition of the unmet need for treatment of anxiety and depression [1] has been accompanied by interest in efficient measurement of the symptoms or distress associated with these disorders [2]. Recent reports confirm that routine monitoring of symptoms and reporting of outcomes can improve the quality and effectiveness of face-to-face psychological care, and reassure agencies funding the care that the services they are paying for are helping the patient [3]. The use of routine monitoring of symptoms and reporting of outcomes has also been applied to the internet-delivery of psychological treatments [4].

The Patient Health Questionnaire 9-Item scale [5] (PHQ-9) and the Generalized Anxiety Disorder 7-item scale [6] (GAD-7) closely follow the DSM-IV diagnostic criteria for depression and generalized anxiety disorder respectively. The GAD-7 is also sensitive to the presence of social phobia, panic disorder and post-traumatic stress disorder [6]. The ten-item Kessler Psychological Distress Scale [7] (K-10) is a well-established screener for general psychological distress and is sensitive to symptoms of both anxiety and depression. They all have well-established benchmarks and are sensitive to change [5–7]. Although the PHQ-9, GAD-7 and K-10 are relatively brief, completing the questionnaires can be time consuming and impractical in some clinical settings. Several studies have shown that the short versions (PHQ-2, GAD-2, and K-6) are reliable screening tools [7–12]. There is less evidence for their usefulness in tracking symptom changes in response to treatment, although at least one study has reported that the PHQ-2 measures symptom severity and outcomes over time [13].

The aims of the current study were to confirm the utility of the PHQ-2, GAD-2, and K-6 as screening instruments and to assess their capacity to measure change in symptoms following treatment. An initial sample was drawn from a series of randomised controlled trials (RCTs) during which the PHQ-9, GAD-7 and K-10 were administered to identify the presence of symptoms and to measure progress. Trial participants were also administered the Mini International Neuropsychiatric Interview

[*] Corresponding author at: eCentreClinic and MindSpot Clinic, Department of Psychology, Macquarie University, Sydney, NSW, Australia.
E-mail address: lauren.staples@mq.edu.au (L.G. Staples).

https://doi.org/10.1016/j.genhosppsych.2018.11.003
Received 6 September 2018; Received in revised form 16 November 2018; Accepted 20 November 2018

(MINI) to establish their diagnoses. Analyses were then replicated where possible using a sample of patients receiving iCBT as part of routine clinical care through the MindSpot Clinic, a national online and telephone mental health clinic funded by the Australian Government. For both groups, participants were enrolled in the Wellbeing course, a transdiagnostic iCBT intervention designed to treat core symptoms of anxiety and depression [14]. Specifically, the hypotheses were that the PHQ-2, GAD-2 and K-6 would identify patients with anxiety and depression and that the short measures would detect treatment changes equivalent to those of the longer versions.

2. Method

2.1. Sample characteristics

Two data samples were used in this study. The first ("research trial sample") came from four related randomised controlled trials that used the same recruitment methodology, and which were conducted to examine the efficacy of internet-delivered treatment for depression and anxiety, described in detail elsewhere [15–18]. Briefly, participants (n = 993) in the RCTs were recruited via the eCentreClinic (www.ecentreclinic.org), a specialist research unit that offers the opportunity for adults with common mental disorders to receive free treatment by participating in research trials. Inclusion criteria determined that participants were residents of Australia aged 18 to 64 years, with at least mild symptoms depression or anxiety. Participants were excluded if they were experiencing unmanaged symptoms of psychosis or very severe depression. They were also excluded if they had self-harmed or made a suicide attempt within the past 12 months, were currently undergoing CBT, or had started or changed psychotropic medication within the past month.

The second data set came from consecutive patients starting the Wellbeing Course at the MindSpot Clinic from 1 January 2017 to 30 June 2017 ("routine care sample"). A total of 1389 patients were eligible for analysis, according to the following criteria: Australian resident eligible for publicly funded health services, 18 years of age or older, reported a principal complaint of anxiety or depression, and gave consent for their de-identified data to be analysed and reported. Individuals who were acutely suicidal, preferred face-to-face services, were currently participating in CBT, or had clinical presentations deemed to require face-to-face assessment (e.g. untreated and disabling psychotic symptoms), were referred to suitable alternative services. Most patients self-referred to the MindSpot Clinic website (www.mindspot.org.au). All patients completed an online assessment, consisting of demographic questions and standardised self-report symptom questionnaires, including the PHQ-9, GAD-7, and K-10. MindSpot patients did not complete a MINI diagnostic assessment. The demographic details and baseline symptoms of both samples are described in Table 1. The RCTs and the analysis of the MindSpot patient data were reviewed and approved by the Human Research Ethics Committee at Macquarie University.

2.2. Measures

2.2.1. Mini International Neuropsychiatric Interview Version 5.0.0 (MINI)

Diagnostic data for depression (Major Depression or Dysthymia) and anxiety (Generalized Anxiety Disorder, Social Anxiety Disorder or Panic Disorder) were obtained during the RCTs via telephone administration of the Mini International Neuropsychiatric Interview Version 5.0.0 (MINI) [19]. The MINI was administered to the research trial sample prior to treatment and again at 3-month follow-up.

2.2.2. Patient Health Questionnaire-9 item (PHQ-9) and 2-item (PHQ-2)

The PHQ-9 consists of nine items measuring symptoms and severity of depression. The PHQ-9 has good internal consistency and is sensitive to change [8]. Each item is rated on a 4-point Likert scale (0 to 3), with increasing scores indicating greater symptom severity. Total scores range from 0 to 27, with a score of 10 or more indicating a diagnosis of depression [5]. The PHQ-2 consists of the first two items of the PHQ-9, which are considered the two core criteria for depressive disorders [20]. These items are: 1.) Feeling down, depressed, or hopeless and 2.) Little interest or pleasure in doing things. Total scores range from 0 to 6. Cut-off scores of ≥3 are indicative of depression on the PHQ-2 [9,20,21].

Table 1
Comparison of patient symptoms and demographics in the two samples.

	Research trial sample [1] (n = 993)	Routine care sample [2] (n = 1389)	Significance
Demographics			
Age (mean and SD)	43.1 years (11.4)	41.1 years (10.0)	$p < .001$ ($F_{1,2380}$ = 20.67)
Gender – female	707 (71.2%)	989 (71.2%)	$p > .05$ (X^2 = 0.00)
Undergraduate or postgraduate degree	544 (54.8%)	678 (48.8%)	$p < .01$ (X^2 = 8.26)
Full or part time employment	728 (73.3%)	977 (70.3%)	$p > .05$ (X^2 = 2.52)
Married or de facto	625 (62.9%)	768 (55.3%)	$p < .001$ (X^2 = 13.95)
Above clinical cut-offs prior to treatment			
PHQ-9 ≥ 10	589 (59.3%)	1034 (74.4%)	$p < .001$ (X^2 = 61.03)
PHQ-2 ≥ 3	460 (46.3%)	881 (63.4%)	$p < .001$ (X^2 = 68.84)
GAD-7 ≥ 8	694 (69.9%)	1114 (80.2%)	$p < .001$ (X^2 = 33.67)
GAD-2 ≥ 3	665 (67.0%)	1064 (76.6%)	$p < .001$ (X^2 = 27.00)
K-10 ≥ 22	682 (68.7%)	1264 (91.0%)	$p < .001$ (X^2 = 192.91)
K-6 ≥ 14	648 (65.3%)	1221 (87.9%)	$p < .001$ (X^2 = 175.77)

[1] eCentreClinic sample.
[2] MindSpot sample.

2.2.3. Generalized Anxiety Disorder Scale 7-item (GAD-7) and 2-item (GAD-2)

The GAD-7 consists of seven items and is sensitive to the presence of generalized anxiety disorder, social phobia, panic disorder and post-traumatic stress disorder [6]. Each item is rated on a 4-point Likert scale (0 to 3), with increasing scores indicating greater symptom severity. Total scores range from 0 to 21, and a score of 8 or more indicates the likely presence of an anxiety disorder [6]. The GAD-2 consists of the first two items of the GAD-7, which are considered core criteria for diagnosing an anxiety disorder [22]. These items are: 1.) Feeling nervous, anxious, or on edge, and 2.) Not being able to stop or control worrying. Total scores range from 0 to 6, with a total score ≥ 3 indicative of a clinically relevant anxiety disorder [10,22,23].

2.2.4. Kessler Psychological Distress Scale 10-item (K-10) and 6-item (K-6)

The K-10 is a nonspecific measure of psychological distress widely used in world mental health surveys [7]. It consists of ten items, rated on a 5-point Likert scale (1–5), with increasing scores indicating greater symptom severity. Total scores range from 10 to 50, with a score ≥ 22 indicative of a clinically relevant mood or anxiety disorder [24]. The K-6 comprises six items of the original scale. These items relate to feelings of nervousness (item 2), hopelessness (item 4), restlessness (item 5), the feeling that everything takes too much effort (item 8), sadness (item 9) and worthlessness (item 10). Total scores on the K-6 range from 6 to 30 [7]. Consistent with guidelines and for the purposes of the current study, a total score ≥ 14 was considered indicative of a clinically relevant anxiety or depression [24,25].

2.3. Statistical analyses

Discriminative validity was assessed using data from the research trial sample. A MINI diagnosis of depression (Major Depressive Disorder or Dysthymia) was used as the reference standard for the PHQ-2, and a MINI diagnosis of anxiety (Generalized Anxiety Disorder, Social Anxiety Disorder, Panic Disorder or Post-Traumatic Stress Disorder) was used as the reference standard for the GAD-2. For the K-6, a MINI diagnosis of depression and/or anxiety was used as the reference standard. For comparison, the discriminative abilities of the longer versions (PHQ-9, GAD-7 and K-10) were also assessed. ROC curve analyses were conducted with results interpreted as acceptable if the area under the curve (AUC) values were between 0.70 and 0.79, and excellent if they were ≥ 0.80 [26]. Criterion group validity was assessed using ANOVA to compare pre-treatment symptom scores on the short-form scales, for patients with and without a diagnosis of depression or anxiety.

Other psychometric properties were analysed using the research trial sample data and replicated where possible using the data from routine care. Test-retest reliability was assessed using the intraclass correlation coefficient (ICC) between initial assessment scores and scores obtained at pre-treatment 1 to 4 weeks later. Based on a 95% confidence interval of the ICC estimate, values of < 0.50 were classified as poor, 0.50–0.75 as moderate, 0.76–0.90 as good, and > 0.90 as excellent. Internal consistency was assessed by calculating Cronbach's alphas at the start of treatment, where an alpha coefficient between 0.70 and 0.79 was considered acceptable, between 0.80 and 0.89 was considered good, and ≥0.90 was considered excellent. Correlations between the short-form scales were also assessed to determine the extent to which they may be measuring similar constructs. Correlation coefficients < 0.30 were considered small, between 0.30 and 0.49 were considered moderate, and correlations of 0.50 or greater were considered large.

To assess responsiveness to treatment effects in both the research trial sample and the routine care sample, generalized estimation equations (GEEs) were used to examine the significance of change from baseline to the follow-up conducted three months after treatment. Only data from participants who completed follow-up were used. Cohen's *d* within-group effect sizes with 95% confidence intervals were also calculated. Sensitivity to change was measured by categorising research trial participants according to their baseline diagnostic status on the MINI (depression, anxiety, both depression and anxiety, or no diagnosis) and comparing means and effect sizes on the short and long forms of the questionnaires. An alpha of $p < .05$ was employed for tests of statistical significance. Statistics were conducted using SPSS version 24.

3. Results

3.1. Sample characteristics

The baseline characteristics of the participants in each sample are compared in Table 1. There were no differences between samples in gender or employment status. The mean age of the routine care sample was lower than the research trial sample (41.1 years compared to 43.1 years). A lower proportion of the routine care sample reported being married or having a tertiary degree. In addition, the proportions of patients reporting symptoms above the clinical thresholds were higher for patients enrolled in routine care.

3.2. Discriminative validity

ROC curve analyses of the short and long-form questionnaires were conducted both prior to treatment and 3-months after treatment (Table 2 and data supplement). At both baseline and follow-up, discriminative validity was excellent for the PHQ-2 and PHQ-9 (AUCs between 0.80 and 0.85, ps < 0.001), acceptable for the GAD-2 and GAD-7 (AUCs between 0.72 and 0.76, ps < 0.001), and acceptable for the K-6 and K-10 (AUCs between 0.76 and 0.78, ps < 0.001).

Sensitivity and specificity at pre-treatment are presented in Table 3. At a clinical cut-off score of ≥3, the PHQ-2 demonstrated the optimal degree of sensitivity and specificity (sensitivity = 0.64, specificity = 0.85; likelihood ratio = 4.19; Youden's J = 0.49), and the GAD-2 also showed optimal sensitivity and adequate specificity at ≥3 (sensitivity = 0.71, specificity = 0.69; likelihood ratio = 2.30; Youden's J = 0.40). At a clinical cut-off score of ≥14, the K-6 demonstrated good sensitivity and specificity (sensitivity = 0.68, specificity = 0.69; likelihood ratio = 2.19; Youden's J = 0.37), however the Youden index was highest at a cut-off of ≥16 (sensitivity = 0.52, specificity = 0.87; likelihood ratio = 4.00; Youden's J = 0.39).

3.3. Other psychometric properties

Test-retest reliability was measured using scores obtained at initial assessment and again before respondents commenced treatment (1 to 4 weeks later). Test-retest reliability in the research sample was good for the PHQ-2 (0.79) and GAD-2 (0.81). Results compared well with the longer measures (PHQ-9 = 0.86; GAD-7 = 0.83). Test-retest data for the K-6 was not available for the research trial sample but was good in the routine care sample (0.88) and compared well with the K-10 (0.89). Test-retest data for the PHQ-2 and GAD-2 was not available for the routine care sample.

Internal consistency in the research sample was good for the PHQ-2 (α = 0.83), GAD-2 (α = 0.81), and K-6 (α = 0.83). All short-forms compared well to their longer versions (PHQ-9: α = 0.85; GAD-7: α = 0.88; K-10: α = 0.88). Internal consistency was replicated in the routine care sample for the PHQ-2 (α = 0.81). Cronbach's alpha coefficients were lower for the GAD-2 and the K-6 in the clinical sample (α = 0.77 and 0.78 respectively) but remained in the acceptable range. The short-forms compared well to the longer versions in the routine care sample (PHQ-9: α = 0.85; GAD-7: α = 0.85; K-10: α = 0.84).

The K-6 demonstrated large correlations with the PHQ-2 in both the research trial sample and the routine care sample (0.70 and 0.71 respectively). The K-6 also showed a large correlation with the GAD-2 in the research trial sample (0.60) and a moderate correlation with the GAD-2 in the routine care sample (0.45). In the research trial sample, the PHQ-2 and GAD-2 were moderately correlated (0.42). In the routine care sample, the correlation between the PHQ-2 and GAD-2 was small (0.28).

3.4. Responsiveness to change

GEE analysis revealed a significant decrease in PHQ-2 scores from baseline to follow-up in both the research trial sample (*Wald's* $\chi^2 = 360.75$, $p < .001$) and the routine care sample (*Wald's* $\chi^2 = 344.90$, $p < .001$). For the GAD-2, there was a significant decrease in scores from baseline to follow-up in the research trial sample (Wald's χ2 = 594.41, $p < .001$) and routine care sample (Wald's χ2 = 481.57, $p < .001$). Similarly, for the K-6, there was a significant decrease in scores from baseline to follow-up in the research trial sample (Wald's χ2 = 657.92, $p < .001$) and routine care sample (Wald's χ2 = 695.85, $p < .001$).

Symptom reductions corresponded to large effect sizes from baseline to follow-up on all measures. Means, standard deviations, Cohen's *d* effect sizes, and percentage changes for each measure are shown in Table 4. For the total research trial sample, percentage change is identical for the PHQ-2 and PHQ-9 (50.0%; CI 46.5%–53.5%) and identical for the GAD-2 and GAD-7 (52.9%; CI 49.4%–56.4%). Percentage changes for the K-6 compared to the K-10 are 46.7% (CI 43.2%–50.2%) and 46.3% (42.8%–49.8%) respectively. Percentage changes were comparable for the short and long versions in the routine care sample and ranged from 48.5% to 55.3%. Table 4 also shows

Table 2
Area under the curve at baseline and follow-up, for patients in the research trial sample with a diagnosis of depression, anxiety, or either (or both) of these.[1]

	Area under the curve (95% CI)					
	Depression		Anxiety		Anxiety and/or depression	
	PHQ-9	PHQ-2	GAD-7	GAD-2	K-10	K-6
Baseline	0.85 (0.82–0.87)	0.83 (0.81–0.86)	0.75 (0.71–0.79)	0.76 (0.72–0.80)	0.78 (0.74–0.83)	0.77 (0.71–0.82)
Follow-up	0.81 (0.77–0.86)	0.80 (0.75–0.85)	0.74 (0.70–0.78)	0.72 (0.69–0.76)	0.77 (0.73–0.80)	0.76 (0.72–0.79)

[1] eCentreClinic sample. AUC: acceptable = 0.70–0.79; excellent = ≥0.80.

Table 3
Sensitivity and specificity of the short-form questionnaires at selected cut-off scores.[1]

Cut-off score	Sensitivity	Specificity	Likelihood ratio	Youden's J
PHQ-2				
≥1	0.99	0.28	1.37	0.27
≥2	0.97	0.48	1.84	0.45
≥3	0.64	0.85	4.19	0.49
≥4	0.44	0.93	6.58	0.37
≥5	0.23	0.98	12.72	0.21
GAD-2				
≥1	0.99	0.13	1.13	0.12
≥2	0.91	0.37	1.44	0.28
≥3	0.71	0.69	2.30	0.40
≥4	0.55	0.81	2.94	0.36
≥5	0.36	0.88	3.09	0.24
K-6				
≥8	0.99	0.15	1.16	0.14
≥9	0.96	0.25	1.28	0.21
≥10	0.93	0.35	1.43	0.28
≥11	0.87	0.43	1.53	0.30
≥12	0.81	0.51	1.65	0.32
≥13	0.75	0.56	1.70	0.31
≥14	0.68	0.69	2.19	0.37
≥15	0.60	0.77	2.61	0.37
≥16	0.52	0.87	4.00	0.39
≥17	0.46	0.89	4.18	0.35
≥18	0.39	0.93	5.57	0.32
≥19	0.33	0.96	8.25	0.29
≥20	0.26	0.96	6.50	0.22
≥21	0.20	0.97	6.67	0.17

[1] eCentreClinic sample. Likelihood ratio = sensitivity/(1-specificity). Common minimum standard = 2.00. Youden's J statistic = sensitivity + specificity – 1. Values range between 0 (low diagnostic accuracy) and 1 (high diagnostic accuracy).

treatment outcomes for the research trial sample grouped by diagnostic status (depression, anxiety, comorbid depression and anxiety, or no diagnosis). Effect sizes and percentage changes were large for all scales regardless of diagnosis.

4. Discussion

The PHQ-2, GAD-2, and K-6 are easily administered via the internet or in person and are brief enough to be administered during most consultations. This study examined their clinical utility in comparison to the longer forms of each questionnaire. The results support the hypotheses that the PHQ-2, GAD-2 and K-6 can identify patients with anxiety and/or depression, and that these measures are sensitive to treatment changes. The diagnostic accuracies of the short versions were comparable to the long versions in all cases. The short versions also showed changes in symptom scores and effect sizes for patients completing treatment and follow-up. In brief, the results support the use of the short form questionnaires to both identify and measure change in symptoms of anxiety and depression.

This study has the advantage of using diagnostic data in conjunction with symptom questionnaires administered to large samples of patients enrolled in research trials of iCBT, as well as those treated with iCBT as part of routine care. The finding that online administration of these short-form questionnaires can identify adults with symptoms of depression and/or anxiety replicates findings using data from patients seen in face-to-face treatment [7,10,11,13,20,22,27]. For both the PHQ-2 and GAD-2, a cut-off score of ≥3 had optimal sensitivity and specificity. For the K-6, the optimal clinical cut-off was more difficult to define. While the Youden's J statistic was highest at a cut-point of ≥16, the difference between sensitivity and specificity was large. At a cut-point of ≥14, the Youden index remained high, and the likelihood ratio remained greater than the common minimum standard. In addition, a cut-point of ≥14 demonstrated a good balance between sensitivity and specificity and is consistent with previous reports [24,25]. The K-6 also demonstrated large correlations with the PHQ-2 and more moderate correlations with the GAD-2. This is perhaps unsurprising given that four of six items measure depressive symptoms (hopelessness, effort, sadness and worthlessness), however factor analysis has shown that the K-6 can be used as a single-factor measure of psychological distress [28,29] and the symptom reductions observed in the current study support its usefulness as a brief measure.

Importantly, each of the short-form questionnaires were responsive to treatment change. Symptom reductions on the short versions corresponded to large within-group effect sizes and percentage changes from baseline to follow-up in both samples of patients. The current results suggest that the use of the PHQ-2, GAD-7, and K-6 in routine monitoring and reporting of treatment outcomes may be practical and effective.

The study has some limitations. First, a diagnostic interview was not administered to the patients enrolled in routine care, limiting the extent to which we can assume the diagnostic validity of the questionnaires in real-world settings. It should also be noted that the clinicians administering the diagnostic interview to research sample participants were not blind to symptom scores obtained at application. Second, despite the very large samples, the use of treatment-seeking patients recruited and contacted online may limit the extent to which the findings might apply in other clinical settings, although the demographic characteristics of MindSpot patients closely match the Australian population [4] and are not unlike patients seen in face to face clinics [30]. Third, the current study did not directly analyse the utility of combining the PHQ-2 and GAD-2 to obtain a single overall "PHQ-4" score as a marker of general psychological distress suggested by other research [9]. Further research to examine the psychometric properties of these scales across cultures, age or other demographic sub-groups would also be useful.

Despite these limitations, the PHQ-2, GAD-2, and K-6 showed good psychometric properties in two heterogeneous samples of treatment-seeking adults. The results were consistent across the two samples, illustrating their usefulness in both research trials and in routine care. Results demonstrate the feasibility of using these very brief measures, either alone or in conjunction with each other, to monitor symptoms and report outcomes in mental health settings and objectively evaluate the effectiveness of mental health care.

Table 4
Means and effect sizes for participants completing follow-up in the research trial sample (analysed by diagnosis) and the routine care sample (diagnostic data not available).

Research trial sample	Baseline mean (SD)	Follow-up mean (SD)	Effect sizes (CI)	Percentage changes
Depression diagnosis (n = 53)				
PHQ-9	11.4 (4.1)	4.9 (5.0)	1.42 (0.99–1.84)	57.0% (43.7%–70.3%)
PHQ-2	3.2 (1.4)	1.2 (1.5)	1.38 (0.95–1.79)	62.5% (49.5%–75.5%)
GAD-7	8.5 (4.6)	3.4 (3.5)	1.25 (0.82–1.65)	60.0% (46.8%–73.2%)
GAD-2	2.6 (1.7)	0.9 (1.2)	1.16 (0.74–1.56)	65.4% (52.6%–78.2%)
K-10	26.2 (5.3)	17.6 (6.5)	1.45 (1.01–1.87)	53.1% (39.7%–66.5%)
K-6	16.3 (3.6)	10.8 (4.3)	1.39 (0.95–1.80)	53.4% (40.0%–66.8%)
Anxiety diagnosis (n = 321)				
PHQ-9	7.1 (4.0)	4.0 (3.3)	0.85 (0.68–1.01)	43.7% (38.3%–49.1%)
PHQ-2	1.5 (1.2)	0.9 (1.1)	0.52 (0.36–68)	40.0% (34.6%–45.4%)
GAD-7	9.3 (4.9)	4.5 (3.6)	1.12 (0.95–1.28)	51.6% (46.1%–57.1%)
GAD-2	3.2 (1.6)	1.6 (1.3)	1.10 (0.93–1.26)	50.0% (44.5%–55.5%)
K-10	21.5 (5.6)	16.7 (5.1)	0.90 (0.73–1.06)	41.7% (36.3%–47.1%)
K-6	13.0 (3.5)	10.2 (3.2)	0.83 (0.67–1.00)	40.0% (34.6%–45.4%)
Both depression and anxiety (n = 355)				
PHQ-9	13.6 (4.7)	6.7 (4.9)	1.44 (1.27–1.60)	50.7% (45.5%–55.9%)
PHQ-2	3.4 (1.4)	1.6 (1.4)	1.29 (1.12–1.45)	52.9% (47.7%–58.1%)
GAD-7	12.5 (4.6)	5.9 (4.7)	1.42 (1.25–1.58)	52.8% (47.6%–58.0%)
GAD-2	4.1 (1.6)	1.9 (1.6)	1.38 (1.21–1.54)	53.7% (48.5%–58.9%)
K-10	29.6 (6.5)	20.2 (7.7)	1.32 (1.16–1.48)	48.0% (42.8%–53.2%)
K-6	18.4 (4.3)	12.4 (4.9)	1.30 (1.14–1.46)	48.4% (43.2%–53.6%)
No diagnosis (n = 61)				
PHQ-9	5.7 (4.4)	3.0 (3.8)	0.66 (0.29–1.02)	47.4% (34.9%–59.9%)
PHQ-2	1.6 (1.4)	0.6 (1.0)	0.82 (0.45–1.19)	62.5% (50.4%–74.6%)
GAD-7	5.4 (3.2)	2.7 (2.8)	0.90 (0.52–1.26)	50.0% (37.5%–62.6%)
GAD-2	1.8 (1.2)	1.0 (1.0)	0.72 (9.35–1.09)	44.4% (31.9%–56.9%)
K-10	18.9 (5.3)	14.4 (4.6)	0.91 (0.53–1.27)	50.6% (38.1%–63.2%)
K-6	11.6 (3.7)	8.7 (2.8)	0.88 (0.51–1.25)	51.8% (39.3%–64.3%)
Total research trial sample (n = 790)				
PHQ-9	**10.2 (5.4)**	**5.1 (4.4)**	**1.04 (0.93–1.14)**	**50.0% (46.5%–53.5%)**
PHQ-2	**2.4 (1.6)**	**1.2 (1.3)**	**0.82 (0.72–0.93)**	**50.0% (46.5%–53.5%)**
GAD-7	**10.4 (5.1)**	**4.9 (4.2)**	**1.18 (1.07–1.28)**	**52.9% (49.4%–56.4%)**
GAD-2	**3.4 (1.7)**	**1.6 (1.4)**	**1.16 (1.05–1.26)**	**52.9% (49.4%–56.4%)**
K-10	**25.2 (7.3)**	**18.1 (6.7)**	**1.01 (0.91–1.12)**	**46.7% (43.2%–50.2%)**
K-6	**15.5 (4.8)**	**11.1 (2.8)**	**1.12 (1.01–1.23)**	**46.3% (42.8%–49.8%)**

Routine care sample	Baseline mean (SD)	Follow-up mean (SD)	Effect sizes (CI)	Percentage changes
Total routine care sample (n = 533)				
PHQ-9	**12.9 (6.0)**	**6.2 (5.3)**	**1.18 (1.05–1.31)**	**51.9% (47.7%–56.1%)**
PHQ-2	**3.2 (1.8)**	**1.5 (1.5)**	**1.03 (0.90–1.15)**	**53.1% (48.9%–57.3%)**
GAD-7	**11.5 (4.9)**	**5.4 (4.4)**	**1.31 (1.18–1.44)**	**53.0% (48.8%–52.2%)**
GAD-2	**3.8 (1.7)**	**1.7 (1.5)**	**1.31 (1.18–1.44)**	**55.3% (51.1%–59.5%)**
K-10	**29.4 (6.6)**	**20.0 (7.0)**	**1.38 (1.25–1.51)**	**48.5% (44.3%–53.7%)**
K-6	**18.3 (4.4)**	**12.3 (4.5)**	**1.35 (1.21–1.48)**	**48.8% (44.6%–53.0%)**

Supplementary data to this article can be found online at https://doi.org/10.1016/j.genhosppsych.2018.11.003.

Acknowledgments

The research trials reported in this paper were funded by the Australian National Health and Medical Research Council Project Grant No. 630560.

References

[1] Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. Bulletin of the. 82. World Health Organization; 2004. p. 858–66.

[2] Batterham PJ, Sunderland M, Carragher N, Calear AL, Mackinnon AJ, Slade T. The Distress Questionnaire-5: population screener for psychological distress was more accurate than the K6/K10. J Clin Epidemiol 2016;71:35–42.

[3] Clark DM, Canvin L, Green J, Layard R, Pilling S, Janecka M. Transparency about the outcomes of mental health services (IAPT approach): an analysis of public data. The Lancet 2018;391:679–86.

[4] Titov N, Dear BF, Staples LG, Bennett-Levy J, Klein B, Rapee RM, et al. The first 30 months of the MindSpot Clinic: evaluation of a national e-mental health service against project objectives. Aust N Z J Psychiatry 2016;51:1227–39.

[5] Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16:606–13.

[6] Spitzer RL, Kroenke K, Williams JB, Lowe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166:1092–7.

[7] Kessler R, Andrews G, Colpe L, Hiripi E, Mroczek D, Normand S, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;32:959–76.

[8] Kroenke K, Spitzer RL, Williams JB, Lowe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010;32:345–59.

[9] Kroenke K, Spitzer RL, Williams JB, Lowe B. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics 2009;50:613–21.

[10] Plummer F, Manea L, Trepel D, McMillan D. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. Gen Hosp Psychiatry 2016;39:24–31.

[11] Furukawa TA, Kessler RC, Slade T, Andrews G. The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. Psychol Med 2003;33:357–62.

[12] Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. BJPsych Open 2016;2(2):127–38.

[13] Lowe B, Kroenke K, Grafe K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). J Psychosom Res 2005;58:163–71.

[14] Titov N, Dear BF, Staples LG, Bennett-Levy J, Klein B, Rapee RM, et al. The first 30 months of the MindSpot Clinic: evaluation of a national e-mental health service against project objectives. Aust N Z J Psychiatry 2017;51:1227–39.

[15] Dear BF, Staples LG, Terides MD, Karin E, Zou J, Johnston L, et al. Transdiagnostic versus disorder-specific and clinician-guided versus self-guided internet-delivered treatment for generalized anxiety disorder and comorbid disorders: a randomized

controlled trial. J Anxiety Disord 2015;36:63–77.

[16] Titov N, Dear BF, Staples LG, Terides MD, Karin E, Sheehan J, et al. Disorder-specific versus transdiagnostic and clinician-guided versus self-guided treatment for major depressive disorder and comorbid anxiety disorders: a randomized controlled trial. J Anxiety Disord 2015;35:88–102.

[17] Dear BF, Staples LG, Terides MD, Fogliati VJ, Sheehan J, Johnston L, et al. Transdiagnostic versus disorder-specific and clinician-guided versus self-guided internet-delivered treatment for Social Anxiety Disorder and comorbid disorders: a randomized controlled trial. J Anxiety Disord 2016;42:30–44.

[18] Fogliati VJ, Dear BF, Staples LG, Terides MD, Sheehan J, Johnston L, et al. Disorder-specific versus transdiagnostic and clinician-guided versus self-guided internet-delivered treatment for panic disorder and comorbid disorders: a randomized controlled trial. J Anxiety Disord 2016;39:88–102.

[19] Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. 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 1998;59(Suppl. 20):22–33.

[20] Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care 2003;41:1284–92.

[21] Richardson LP, Rockhill C, Russo JE, Grossman DC, Richards J, McCarty C, et al. Evaluation of the PHQ-2 as a brief screen for detecting major depression among adolescents. Pediatrics 2010;125:e1097–103.

[22] Kroenke K, Spitzer RL, Williams JB, Monahan PO, Lowe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med 2007;146:317–25.

[23] Christensen H, Batterham PJ, Grant JB, Griffiths KM, Mackinnon AJ. A population study comparing screening performance of prototypes for depression and anxiety with standard scales. BMC Med Res Methodol 2011;11:154.

[24] Australian Bureau of Statistics. Information paper: use of the Kessler Psychological Distress Scale in ABS health surveys. Australia, 2007-08 Canberra: Australian Bureau of Statistics; 2012.

[25] Cornelius BL, Groothoff JW, van der Klink JJ, Brouwer S. The performance of the K10, K6 and GHQ-12 to screen for present state DSM-IV disorders among disability claimants. BMC Public Health 2013;13:128.

[26] Hosmer D, Lemeshow S. Applied logistic regression. New York: John Wiley and Sons; 2000.

[27] Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, Fishman T, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. Ann Fam Med 2010;8:348–53.

[28] Ko J, Harrington D. Factor structure and validity of the K6 scale for adults with suicidal ideation. J Soc Soc Work Res 2016;7:43–63.

[29] Brooks RT, Beard J, Steel Z. Factor structure and interpretation of the K10. Psychol Assess 2006;18:62–70.

[30] Titov N, Andrews G, Kemp A, Robinson E. Characteristics of adults with anxiety or depression treated at an internet clinic: comparison with a national survey and an outpatient clinic. PLoS One 2010;5:e10885.