Undergraduate Research Journal

Volume 25 Article 8

2021

Assessment Tools in Clinical and Counseling Psychology

Alea Reifenrath
University of Nebraska at Kearney

ASSESSMENT TOOLS IN CLINICAL AND COUNSELING PSYCHOLOGY

Alea Reifenrath
Mentor: Dr. Chris Waples, Department of Psychology

OVERVIEW

"Evidence-based" has become a keyword in the realm of clinical and counseling psychology. Hundreds of thousands of articles have been written on the topic. It represents a focus on empirical evidence and best practice. This key phrase can be tied to many facets in psychology and health care, such as evidence-based practice and evidence-based treatment. One area which is often overlooked is evidence-based assessment (EBA). EBA is defined as the use of standardized assessment tools that have research support for their psychometric properties (Jensen-Doss & Hawley, 2010). Assessment is a key aspect of the therapeutic process. It involves screening individuals who are at risk for a particular mental health concern, determining the severity of a given issue, and monitoring the effectiveness of treatment. EBA allows for best practice by providing empirically sound, valid, reliable measures of assessment. Jensen-Doss and Hawley (2010) point out that assessment is crucial to treatment, as all treatment is based on assessment and diagnosis. Therefore, they claim, some element of EBA must be included for evidence-based treatment to be fully effective and accurate.

Despite the potential benefits of EBA, it has received little attention. While there is a trend toward evidence-based practice and evidence-based treatment, EBA seems to remain unnoticed (Beidas et al., 2015). As many as 83% of psychiatrists report never using standardized assessment tools (Jensen-Doss & Hawley, 2010). Greater use can be seen among psychologists and mental health practitioners, but the majority still report never using such instruments.

A myriad of reasons can be cited for not using EBA tools. Many practical concerns exist, such as limited access to tests, high paperwork burdens, and intense time demands (Beidas et al., 2015). These concerns do make sense, particularly in understaffed environments and small private practice settings. Assessment tools also face skepticism in their ability to provide new information, which could not be obtained through other methods such as interviews. In one survey clinicians reported doubt that such instruments would provide a benefit over clinical judgment (Jensen-Doss & Hawley, 2010). Some clinicians also reported skepticism of the reliability and validity of assessment measures. Overall, however, the most negative views of standardized assessment revolved around practicality. Practicality concerns were found to be the only predictor of assessment tools use.

Should such measures be written off as impractical? To establish that, one must weigh the cost and benefits of assessment tools. These will vary depending on the measure, but let us first examine the big picture. Unstructured or semi-structured interviews remain the most common assessment tool (Jensen-Doss & Hawley, 2010). They are flexible, and allow for in-depth and personalized assessment. However, they have low validity and are prone to bias. Adding standardized assessment measures to unstructured or semi-structured interviews could provide a more well-rounded view. The two complement each other well. Assessment tools would add an element of empiricism while reducing bias; the interview would add an element of client-centered discussion while filling in any gaps of standardized assessment.

Standardized assessment tools also provide a method for reducing clinician bias. Beidas et al. (2015) describe the assessment process as "inherently a decision-making task fraught with biases (p. 2, lines 8-9)". Even experienced and competent clinicians are subject to confirmatory bias and cognitive heuristics. EBA, however, could help to minimize these issues. When used for progress tracking, EBA can also provide feedback for clinicians. It can aid in the evaluation of treatment and allow for best practice. The APA is currently examining the utility of evidence-based assessment tools in a shift towards a dimensional view of disorders (Bastiaens & Galus, 2017). A handful of measures were included in the Diagnostic and Statistical Manuel-5 (DSM-5), and listed as "emerging measures". Such measures could be used to assess the severity of symptoms for a wide array of disorders.

Clearly, EBA has a lot to offer. Yet, as previously discussed, practical concerns can prevent assessment tools from being used. Beidas et al. (2015) list a number of factors to consider, which affect the practicality of an instrument. They state that any assessment tool must be "brief, low cost, valid, reliable, applicable, useful, actionable, and straightforward to administer, score and interpret" (Beidas et al., 2015, p. 3). While this is a rather long list, it does provide many important factors to look for. Instruments that meet this description would provide beneficial information with relatively few drawbacks. A wide array of assessment tools is available, many of which do fit these factors and many of which do not. Beidas et al. report throwing out over 100 instruments in their study, which did not meet the above factors. For this paper, three instruments were chosen based on their popularity and applicability. Beck's Depression Inventory-II, the Patient Health Questionnaire-9, and the Generalized Anxiety Disorder 7-item scale are three of the most used assessment tools. They represent common disorders—depression and anxiety—making them widely applicable. They have also been used in a variety of settings, such as in-patient units, out-patient clinics, and primary care offices. In order to assess their usefulness as evidence-based assessment tools, it is important that their psychometric properties be analyzed.

BECK'S DEPRESSION INVENTORY-II

Beck's Depression Inventory-II (BDI-II) is a highly popular example of an assessment tool. It is a self-report measure of depressive symptoms for individuals age 13 and up (Segal et al., 2008). It screens for depressive symptoms present over the last 2 weeks. The items were modeled after the Diagnostic and Statistical Manuel of Mental Disorders fourth edition (DSM-IV) criteria for a major depressive episode such as depressed mood, anhedonia, low energy, and changes in weight or sleep (Wang & Gorenstein, 2013). The items in the BDI-II are also consistent with the updated version of the DSM, the DSM-5, as the criteria for major depressive episodes were not changed (American Psychiatric Association [APA], 2013). The measure consists of 21 items, each with a 4-point Likert-type scale. Higher scores indicate higher levels of depression.

Overall, the BDI-II has been found to show excellent reliability. A meta-analysis revealed internal consistency averaging around 0.90, which is considered excellent (Wang & Gorenstein, 2013). It was found to range moderately but remained very good to excellent in all samples. Segal et al. (2008) found that the internal consistency was slightly higher for younger adults compared to older adults, but that it was very good to excellent in both cases. The individual items which had the strongest correlation with the total scores were related to loss of interest, loss of pleasure, and sadness. Test-retest reliability was also stable. Coefficients ranged from 0.73-0.96, which is

considered excellent (Wang & Gorenstein, 2013). Longer intervals between applications were associated with smaller correlations between test scores. This is expected, because depressive symptoms can decrease over time even without intervention. There is also some evidence that decreases in retest scores are due to measurement effects (Wang & Gorenstein, 2013). This suggests that continued re-application of the BDI-II in healthcare settings may underestimate the level of depressive symptoms. The measurement effect appears to be relatively small, but it is worth noting. Overall the BDI-II has excellent internal consistency and test-retest reliability.

Construct validity is another important factor to consider. It refers to the ability of a test to measure the concept that it is intended to measure. The construct validity of the BDI-II is typically assessed using confirmatory factor analysis. Some debate remains over whether a two-factor or a single factor structure is a better fit for the data. Some literature suggests that the BDI-II taps into two factors, cognitive-affective and somatic-vegetative (Wang & Gorenstein, 2013). Cognitive-affective items include statements about self-criticism, irritability, guilt, and hopelessness. Somatic-vegetative items include statements about changes in appetite, difficulty concentrating, feelings of restlessness, and loss of energy. This was the factor structure suggested by the original authors, and it has been well supported. However, some studies have found that the two-factor model was not a good fit using confirmatory factor analysis (Segal et al., 2008). 20 out of the 21 items were found to load saliently onto a single factor, suggesting that there is no need to make a distinction between cognitive and somatic items. Segal et al. (2008) claimed that the two factors were not "conceptually meaningful or practically useful" (p. 15). If the two factors are indeed distinct, they appear to be highly related, with between-factor correlation ranging from 0.49 to 0.87 (Wang & Gorenstein, 2013). Thus the cognitive-affective and somatic-vegetative aspects of the BDI-II are closely related, and possibly tap into the same factor.

Convergent validity for BDI-II has been well established (Segal et al., 2008). It has relatively strong, positive correlations with other measures of depression, including the Center for Epidemiologic Studies Depression Scale (CES-D), the Coolidge Axis II Inventory (CATI) depression subscale, the Depression Anxiety Stress Scales depression subscale (DASS-D), and the Symptom Checklist-90 depression subscale (SCL-90-D; Segal et al., 2008; Wang & Gorenstein, 2013). It also has strong positive correlations with measures of general psychopathology and negative correlations with measures of psychological well-being (Wang & Gorenstein, 2013). Correlations with measures of anxiety are also quite strong, ranging from 0.53 to 0.67 (Segal et al., 2008). This suggests the BDI-II offers poor discriminant validity between depressive symptoms and anxiety-based symptoms. Given the overlap in symptoms and the high rates of comorbidity (Wang & Gorenstein, 2013), it is unsurprising that a short assessment tool would be unequipped for distinguishing between the two. However, these strong correlations with anxiety and general psychopathology suggest that the BDI-II should not be used as a sole indicator of depression. Other disorders, physical illness, or social problems could also cause a high score. The measure's high face validity also brings the possibility of malingering (Segal et al., 2008). So while the BDI-II clearly has strong convergent validity with other measures of depression, a more thorough assessment is needed to establish if a diagnosis of depression is merited.

Some divergent validity has been established as well. Divergent validity refers to the ability of the test to avoid being influenced by concepts that it was not intended to measure. Low correlations have been found between the BDI-II and measures of drug and alcohol use (Wang &

Gorenstein, 2013). This an important type of divergent validity to establish, because the DSM-5 criteria for a depressive episode require that it not be due to the physiological effects of a substance or substance abuse (APA, 2013). It is important that any assessment tool of depression does not inadvertently measure drug or alcohol use.

Criterion validity examines how well a test correlates with a target criterion. For the BDI-II, criterion validity assesses how well it predicts a diagnosis of depression. The sensitivity of the measure was found to be 0.70 when compared with structured intake interviews. Wang & Gorenstein (2013) consider this the most important indicator, because it suggests a low level of false negatives, or undetected cases of depression. False negatives are more detrimental than false positives because the suffering individual is unlikely to get help if the depressive symptoms go unrecognized. False positives are unhelpful, but they can be ruled out after further examination. The area under the ROC curve, which indicates diagnostic accuracy was 75% or higher, with an average of 87%. This suggests that the BDI-II has an acceptable to excellent ability to discriminate between individuals with depression and individuals without depression.

Wang and Gorenstein (2013) also suggest having different criterion-referenced cutoff points for different populations. They found that a score of 22 or higher was most predictive in psychiatric samples, while cutoffs of 10 and 15 were most predictive in non-clinical and medical samples, respectively. This is in line with the original authors' recommendation to develop local norms for the interpretation of scores. Psychiatric samples had significantly higher mean scores than medical or nonclinical samples, which is expected given that psychiatric samples are definitionally experiencing higher levels of psychiatric distress. Higher criterion-referenced cutoff points were needed to enhance specificity. The cutoff that should be used for intervention should depend, therefore, on the setting in which it is being used.

Overall, Beck's Depression Inventory provides a short, simple, reliable, and validated measure of depressive symptoms. It has excellent internal consistency and test-retest reliability, it correlates strongly with other measures of depression, and it is able to predict diagnoses of depression in both psychiatric and nonpsychiatric samples. There is a unanimous consensus that the BDI-II works well as a screening tool, and should be accompanied by other assessment methods, such as intake interviews (Segal et al., 2008). The BDI-II can provide psychometrically sound information on depressive symptoms, which is useful for the detection of cases, treatment planning, and progress monitoring.

One of the major disadvantages of the BDI-II is that it is copyrighted and must be obtained from the publisher, which prevents more widespread use (Wang & Gorenstein, 2013). A plethora of other measures of depression is available for free, such as the Patient Health Questionnaire depression module (PHQ-9). But does this measure hold up compared to the BDI-II? Is it psychometrically sound?

PATIENT HEALTH QUESTIONNAIRE-9

The PHQ-9 was developed to screen for depression in primary care settings (Gilbody et al., 2007). It consists of 9 self-report items that are based on the diagnostic criteria for a major depressive episode as defined in the DSM-IV. It is a Likert-type scale, with items scored 0-3. PHQ-

9 refers to the depression module of the Patient Health Questionnaire. The questionnaire as a whole consists of 59 items that screen for a variety of mental health issues, such as alcohol abuse, bulimia nervosa, anxiety, and panic attacks (Spritzer et al., 1999). The modules can be used separately or as a whole, depending on the purpose of the assessment and the available time to complete it. The PHQ-9 can provide a measure of the presence and severity of depressive symptoms. The instruction manual recommends treating scores of 10 or more as a "yellow flag", indicating moderate levels of depressive symptoms and a potentially clinically significant condition. A score of 15 or more represents a "red flag", indicating high levels of depression that should receive immediate assessment and possibly treatment. Moderate and high scores should prompt further assessment, including a clinical interview.

The reliability of the PHQ-9 is well established. Internal consistency of the measure has been found to be around 0.87, which is considered very good (Beard et al., 2016). The test-retest reliability was found to 0.78 for a 2-week interval. This decreased with longer intervals, but as with the BDI-II, such a decrease is expected. Depressive symptoms tend to improve over time, so this decrease is likely due to the instrument picking up on a reduction of symptoms. Overall. it appears to be a stable measure.

The content validity of the PHQ-9 has been assessed using confirmatory factor analysis. It appears to have a two-factor structure, with one factor tapping into cognitive/affective traits of depression and the other tapping into somatic traits (Beard et al., 2016). The cognitive/affective factor includes items that relate to anhedonia, feelings of hopelessness, and feelings of failure. The somatic factor includes items that relate to change in sleep, changes in appetite, and restlessness. Similar to BDI-II, there is some debate over whether these represent two separate factors or whether they load onto the same overarching factor. The two-factor model accounts for around 60% of the variance.

The construct validity of the PHQ-9 has also been established. It correlates strongly with other well-established measures of depression, such as BDI-II and the Center for Epidemiologic Studies Depression Scale (CES-D), with correlations around 0.80 (Beard et al., 2016). The PHQ-9 also correlates strongly with measures of anxiety, such as the anxiety scale of the PHQ. Pearson's *r* values are around 0.60. As with BDI-II, high correlations between measures of depression and anxiety are unsurprising given the high comorbidity rates and the overlap of symptoms. This high correlation is therefore not particularly concerning, although it is important to note that individuals with high levels of anxiety may score highly on the PHQ-9 because of it. Questionnaire administrators should be aware that a moderate or high score on the PHQ-9 could result from mental health issues other than depression.

Criterion validity has been established in multiple ways. Patients in a behavioral health unit, who were experiencing current major depressive episodes, scored significantly higher than patients who were not experiencing current major depressive episodes (Beard et al., 2016). This suggests that the PHQ-9 was able to predict current major depressive episodes, which provides concurrent predictive validity for the measure. The sensitivity and specificity of the measure have also been examined. A meta-analysis found the average sensitivity in a variety of settings to be 0.92 (Gilbody et al., 2007). In other words, the measure accurately detects cases of existing depressive symptoms in 92% of cases. This is quite a bit higher than the specificity of the BDI-II,

which was found to have a sensitivity of around 0.70 (Wang & Gorenstein, 2013). The PHQ-9 was more likely to detect existing depressive symptoms than BDI-II. The average specificity of the PHQ-9 was found to be around 0.80 (Gilbody et al., 2007). This means that the PHQ-9 will correctly reject the presence of depressive symptoms 80% of the time, while falsely identifying an individual as having depressive symptoms 20% of the time. High sensitivity is considered more important than high specificity in this situation, because further evaluation can reduce the number of false positives. Undetected cases of depression, however, can bring about high levels of distress, dysfunction, and danger to an individual. The PHQ-9 displays exceptionally high levels of both specificity and sensitivity, which make it an attractive measure.

It is also sensitive to change, which is important in long-term settings. Significant pre-post changes in scores were found upon discharge from an in-patient behavioral health hospital (Beard et al., 2016). Large effect sizes were observable after only 1-2 weeks of Cognitive-Behavioral Therapy. These effect sizes were found to be similar to the changes observable in CES-D scores. This suggests that the PHQ-9 is able to detect reductions in depressive symptoms. Self-report measures of treatment reflection also support this. Patients with significant reductions in PHQ-9 scores at discharge were more likely to report improvement than patients without significantly lower scores. Therefore, the PHQ-9 is a useful measure of symptom severity and improvement due to treatment. It can be used to assess treatment in psychiatric settings.

The PHQ-9 provides a valid and reliable measure of depressive symptoms. It is also exceptionally brief and easy to use. It was originally designed to screen for depression in primary care settings, although it has been found to be similarly useful in psychiatric and hospital settings (Gilbody et al., 2007). It can be used to detect depression, assess severity, aid in initial treatment decisions, and measure treatment outcome (Spritzer et al., 1999). Its brevity and accessibility, along with its solid psychometrics, allow the PHQ-9 to be widely used. The PHQ-9 is so well regarded that it is listed in the DSM-5 as a recommended measure of depression severity (APA, 2013). It is important to note that the PHQ-9 alone should not be used to diagnose depression. It should only be used to supplement other aspects of assessment. However, it is a valuable evidence-based addition for screening or treating depression in a variety of settings.

GENERALIZED ANXIETY DISORDER 7-ITEM SCALE

Other similar measures exist for common diagnoses, such as anxiety disorders. The Generalized Anxiety Disorder 7-item scale (GAD-7) was created in conjunction with the PHQ-9 (Spritzer et al., 2006). It is extremely brief, consisting of just 7 self-report items on a 4-point Likert-type scale (Beard et al., 2016). Similar to the PHQ-9, the items are based on DSM-IV criteria. They focus on symptoms related to Generalized Anxiety Disorder (GAD), such as restlessness, irritability, and uncontrolled worry (Spritzer et al., 2006). The items are consistent with the latest version of the DSM, although they are slightly more vague than the DSM-5 criteria. For example, one item on the GAD-7 mentions worrying about "different things" while the DSM-5 specifically mentions the worry must occur in multiple facets of life (APA, 2013). The GAD-7 also has a different time frame. It focuses on symptoms over the last two weeks, rather than the 6 months mandated by the DSM-5. This suggests that individuals may score highly on the assessment but not qualify for a diagnosis of GAD.

However, high scores could indicate other mental health issues, particularly other anxiety disorders. For example, an individual who has been experiencing extreme worry for the past month would likely score highly on the GAD-7. They would not qualify for a diagnosis of GAD, but they could potentially be diagnosed with an unspecified anxiety disorder or other specified anxiety disorder. Individuals suffering from a specific phobia or social anxiety disorder would similarly be likely to score highly on the GAD-7 because it does not examine the source or setting of the symptoms. So while the GAD-7 does not discriminate well between the various anxiety disorders, it is well supported as a transdiagnostic measure of anxiety (Beard & Bjorgvinsson, 2014).

The reliability of the measure appears to be reasonable. Internal consistency was found to be 0.88, which is considered very good (Beard & Bjorgvinsson, 2014). Other reliability estimates were found to be around 0.85 (Rutter & Brown, 2017). High reliability estimates were found even when presented in different languages and settings (such as primary care, psychiatric inpatient, and outpatient). Overall, its stability has been well-established.

Construct validity of the GAD-7 has been examined using confirmatory factor analysis. A one-factor model did not fit the data well (Beard & Bjorgvinsson, 2014; Rutter & Brown, 2017). Somatic items, such as difficulty relaxing and restlessness, were found to load onto a separate factor. Rutter and Brown (2017), however, proposed a revised one-factor model that fit the data well. Factor loadings were consistent across gender and race.

Convergent validity for the GAD-7 has been well documented. It has been found to have strong, positive correlations with established measures of stress, anxiety, and worry (such as the Depression Anxiety and Stress Scales and the Penn State Worry questionnaire; Rutter & Brown, 2017). It also had a moderate negative correlation with psychological well-being and a moderate positive correlation with depression (Beard & Bjorgvinsson, 2014). Rutter and Brown (2017) examined measures of depression and obsessive-compulsive disorder (OCD) for discriminate validity. They found moderate correlations with the GAD-7 for each. However, these disorders are highly related to GAD and it is fairly unsurprising that a moderate correlation would exist. Rutter and Brown chose measures of depression and OCD to examine the GAD-7's ability (or lack thereof) to discriminate between disorders. They do, however, point out that the disorders are phenotypically similar and frequently comorbid. A strong correlation with measures of anxiety and worry, as well as moderate correlations with measures of OCD and depression, suggest that the GAD-7 has some ability to differentiate between disorders but that it is easily influenced by other mental health issues.

This weakness in the GAD-7 can also be seen when examining the specificity and sensitivity of the measure. The recommended cutoff score for further evaluation is 10 out of 21. This cutoff was found to be reasonable for nonpsychiatric samples but resulted in high false-positive rates in clinical psychiatric samples (Beard & Bjorgvinsson, 2014). A cutoff score of 10 resulted in fair sensitivity, but poor specificity in psychiatric samples. Considering the previously mentioned correlations with depression and OCD, it is likely that the GAD-7 is influenced by other disorders, which results in artificially high indicators of anxiety. It is also worth noting that individuals entering an inpatient program are likely to be experiencing high levels of stress and anxiety related to the situation and setting. They may be more likely to report that they have "felt

nervous" or "had trouble relaxing" over the last few days because the situation itself could be creating stress and anxiety.

Nonetheless, the GAD-7 has the worst specificity of any measure listed in this paper. Rutter and Brown (2017) claimed that no cutoff score was able to adequately balance sensitivity and specificity. It is clear that, at least in psychiatric settings, the GAD-7 performs poorly as a screening tool for GAD. It lacks the ability to adequately distinguish individuals with GAD from those without. It has been found to perform much better as a screener in primary care settings, which is what it was originally created for (Beard & Bjorgvinsson, 2014).

While the GAD-7 performs poorly as a screener in psychiatric settings, it has had success as a measure of anxiety severity. Pre-post changes with large effect sizes were evident in individuals with GAD, post-traumatic stress disorder, and panic disorder following inpatient treatment (Beard & Bjorgvinsson, 2014). Patients who reported "very much" improvement also had significantly larger reductions in scores than patients who did not report high levels of improvement. Scores decreased an average of 4 points following inpatient treatment, which shows that the GAD-7 is sensitive to change even over a short period of time (Rutter & Brown, 2017). The GAD-7 appears to be a psychometrically sound measure of the severity of anxiety in a broad sense, although it is less successful at screening for the presence or absence of GAD.

Like the PHQ-9, the GAD-7 is short, easy to score, and publicly available. This can make it attractive as a quick and easy addition in a variety of settings. It is also psychometrically sound. It has very good reliability, and evidence of construct and convergent validity. Its uses vary somewhat depending on the setting. In primary care settings it can provide a quick and efficient screening for anxiety disorders (Rutter & Brown, 2017). Because anxiety is a relatively common mental health concern, it can be helpful to assess and screen for it in the general population. In outpatient or inpatient psychiatric settings, the GAD-7 should not be used as a screening tool, as it fails to discriminate well between disorders. However, individuals in these settings have already been identified as needing treatment and will be going through a more thorough assessment to establish a diagnosis. Screening tools, therefore, are less necessary. The GAD-7 does provide a measure of anxiety severity, which can be informative at the onset of treatment to establish which route to take, as well as throughout treatment, to establish the effectiveness of the treatment. Although the GAD-7 has some weaknesses, it is a well-supported assessment tool that can provide supplemental information in a variety of settings.

CONCLUSION

The BDI-II, PHQ-9, and GAD-7 are three examples of a myriad of available assessment tools. The BDI-II is a highly popular instrument, with strong psychometrics. The PHQ-9 provides an alternative to the BDI-II, which is free to the public and still psychometrically sound. The GAD-7 was made in conjunction with the PHQ-9. While it has some weaknesses, it provides a reliable and valid measure of anxiety. These three instruments were chosen due to their popularity, ease of administration, and potential for even greater use. The broad areas of anxiety and depression represent extremely common disorders. Because of their high prevalence, evidence-based assessment tools catered toward depression and anxiety could be useful in primary care offices, inpatient clinics, and outpatient facilities. These measures take only minutes to complete and are

easy to score and interpret. This reduces the strain on time and resources that longer, more complex measures would create. This simplicity and brevity, along with solid psychometric properties, make these measures practical additions to the assessment process. Thus, they address the main concerns of practitioners—that the measures be valid, reliable, practical, and useful. These instruments, along with many others, show that EBA is possible and plausible. The term "evidence-based" should not be reserved for treatment alone, but rather it should be integrated into all parts of the therapeutic process.

REFERENCES

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Bastiaens, L., & Galus, J. (2017). The DSM-5 self-rated level 1 cross-cutting symptom measure as a screening tool. *Psychiatric Quarterly*, *89*(1), 111-115. https://.doi.org/10.1007/s11126-017-9518-7

Beard, C., Hsu, K. J., Rifkin, L. S., Busch, A. B., & Björgvinsson, T. (2016). Validation of the PHQ-9 in a psychiatric sample. *Journal of Affective Disorders*, *193*, 267-273. http://dx.doi.org/10.1016/j.jad.2015.12.075

Beard, C., & Björgvinsson, T. (2014). Beyond generalized anxiety disorder: Psychometric properties of the GAD-7 in a heterogeneous psychiatric sample. *Journal of Anxiety Disorders*, *28*, 547-552. http://dx.doi.org/10.1016/j.janxdis.2014.06.002

Beidas, R. S., Stewart, R. E., Walsh, L., Lucas, S., Downey, M. M., Jackson, K., Fernandez, T., & Mandell, D. S. (2015). Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive and Behavioral Practice*, *22*(1), 5-19. https://doi.org/10.1016/j.cbpra.2014.02.002

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. *Journal of General Internal Medicine*, *22*(11), 1596-1602. https://doi.org/10.1007/s11606-007-0333-y

Jensen-Doss, A. & Hawley, K. M. (2010). Understanding barriers to evidence-based assessment: Clinician attitudes toward standardized assessment tools. *Journal of Clinical Child and Adolescent Psychology*, *39*(6), 885-896. https://doi.org/10.1080/15374416.2010.517169

Rutter, L. A. & Brown, T. A. (2017). Psychometric properties of the Generalized Anxiety Disorder Scale-7 (GAD-7) in outpatients with anxiety and mood disorders. *Journal of Psychopathology and Behavioral Assessment*, *39*(1), 140-146. http://doi.org/10.1007/s10862-016-9571-9

Segal, D. L., Coolidge, F. L., Cahill, B. S., & O'Riley, A. A. (2008). Psychometric properties of the Beck Depression Inventory-II (BDI-II) amid community-dwelling older adults. *Behavior Modification*, *32*(1), 3-20. http://doi.org/10.1177/0145445507303833

Spritzer, R. L., Kroenke, K., Williams, J. W., & Lowe, B. (1999). *Patient Health Questionnaire (PHQ-9): Technical manual.* https://www.phqscreeners.com/

Spritzer, R. L., Kroenke, K., Williams, J. W., & Lowe, B. (2006). *Generalized Anxiety Disorder 7-item Scale (GAD-7): Technical manual.* https://www.phqscreeners.com/

Wang, Y.-P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: A comprehensive review. *Revista Brasileira de Psiquiatria*, *35*, 416-431. http://doi.org/10.1590/1516-4446-2012-1048