CALIFORNIA STATE UNIVERSITY SAN MARCOS

THESIS SIGNATURE PAGE

THESIS SUBMITTED IN PARTIAL FULLFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE

MASTER OF ARTS

IN

PSYCHOLOGY

THESIS TITLE: Evaluating a Computerized Depression Screening Tool with an American
Indian and Alaska Native Sample

AUTHOR: Juan A. Luna

DATE OF SUCCESSFUL DEFENSE: 12/2/13

THE THESIS HAS BEEN ACCEPTED BY THE THESIS COMMITTEE IN
PARTIAL FULLFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ARTS
IN PSYCHOLOGY.

Dr. Gerardo M. Gonzalez
THESIS COMMITTEE CHAIR

[Signature]

12/2/13

Dr. Marie Thomas
THESIS COMMITTEE MEMBER

[Signature]

12/2/13

Dr. Kimberley Pulvers
THESIS COMMITTEE MEMBER

[Signature]

12/2/13
Evaluating a Computerized Depression Screening Tool with an American Indian and Alaska Native Sample

Juan Antonio Luna

California State University, San Marcos
Abstract

Major depression is a significant health concern affecting approximately 14.8 million adults in the United States each year and is commonly undetected in primary care settings. This study evaluated the use of the Voice-Interactive Depression Assessment System (VIDAS) with a sample of 100 (70 females and 30 males) American Indians and Alaska Natives adults from a primary care organization in a rural area of California. The following depression measures were administered in random order: VIDAS, the paper-and-pencil version of the Center for Epidemiological Studies-Depression Scale (CES-D), the paper-and-pencil version of the Beck Depression Inventory-II (BDI-II), the Patient Health Questionnaire (PHQ-9); followed by the Computer Aversion Scale (CAVS) and the Preference for Mode of Administration Questionnaire (PMAQ). The results revealed that VIDAS demonstrated equivalence of test formats (computerized and paper-and-pencil), strong inter-item reliability, and convergent validity. In addition, VIDAS was generally acceptable to participants, and participants did not have a test format preference. These findings suggest that a computerized voice interactive depression screening tool is a viable alternative to traditional screening tools in a rural health care setting. Offering a psychometrically sound assessment alternative that clients find acceptable can aid primary care providers in detecting clients who may be experiencing depression. Overall, the implications for the study are promising and can lead to the further development of appropriate screening tools.
Depression Prevalence: Review of Detection and Screening Tools

Major depression is a significant health concern affecting approximately 14.8 million or 6.7% of adults each year in the United States (Kessler, Chiu, Demier, & Walters, 2005). The Surgeon General (1999) equates major depression to blindness or paraplegia in terms of burden of disability. Unipolar depression is the leading cause of years of healthy life lost as a disability, at 11% for high-income countries and 9.1% for low-and middle-income countries (Lopez, Mathers, Ezzati, Jamison, & Murray, 2006). Furthermore, unipolar depression accounts for 5.6% total disability-adjusted life years among high-income countries (Lopez et al., 2006).

Surprisingly, many of those who suffer from depression may not know they have it, nor are their symptoms detected by their primary care doctor. In fact, one study reported that the physicians who were assessed detected only 28% of depressive disorders in their patients (Coyne, Schwenk, & Fechner-Bates, 1995). Another study found that 32% of primary care patients who had major depression remained undetected for up to a year (Rost et al., 1998). Likewise, Parker et al. (1997) observed undetected mood disorders in 15 patients of the 100 American Indians sampled. A more recent study with military veterans in a primary care setting revealed high undetection of severe depressive symptoms at 36% of the total sample (Liu, Campbell, Chaney, Li, McDonell, Fihn, 2006). Although research reveals a need for better depression detection methods or tools, not all studies indicate poor detection. For instance, one study found primary care physicians given mental health assessment training detected depression more effectively, and had a 76% agreement with mental health clinicians on the diagnoses (Miller & McCrone, 2005). Such results are promising but more research into improving detection methods is needed because this sample only included male military veterans.
Problems associated with nondetection and underdetection is a major concern because so many individuals use primary care for their health care needs. In fact, a majority (72%) of the U.S. population has visited a primary care facility (Krauss, Machlin, & Kass, 1999), and a significant amount (36.5%) of all American Indians and Alaska Natives (AI/ANs) utilize Indian Health Service (IHS) for their primary health care needs (Ogunwale, 2002). IHS supplies primary care services free of charge to AI/ANs enrolled as tribal members (Zuckerman, Haley, Roubideaux, & Lillie-Blanton, 2004). Considering many AI/ANs use primary care facilities through IHS, efficient screening tools are needed in aiding in detecting individuals at risk for developing serious mental health problems.

Issues related to mental health in AI/AN populations are not well understood because there is little research in this area. For instance, AI/ANs are not listed in any of the national mental health epidemiological studies, which include the Epidemiological Catchment Area Study (Robins & Regier, 1991), the National Comorbidity Survey (Kessler, 1994), or the National Comorbidity Survey-Replication (Kessler et al., 2003; Kessler et al., 2005; Kessler et al., 2005). Research with AI/AN populations are often not conducted because of methodological issues such as not having large enough sample sizes, not having enough funding to sustain research, racial misclassification of the population of interest, and culturally inappropriate instrumentation (Duran et al., 2004).

The few studies conducted have mixed results with some revealing a higher incidence of mental health issues for AI/ANs than the general population and others revealing the contrary. One study illustrating a higher incidence of major depression was observed in a sample of 234 women from a primary care facility in Albuquerque, New Mexico (Duran et al., 2004). The prevalence of major depression for women sampled was 18.8% for the past year, and 41.5% of
prevalence over their lifetime. One can see the higher level of depression for this sample when compared to national estimates for the general population at 6.7% for the past year (Kessler et al., 2005) and 16.6% for lifetime prevalence (Kessler et al., 2005). Another study (Parker et al., 1997), observed the prevalence of major depression to be 10% of the total sample (66 females and 34 males; mean age = 37.7), which is also higher than the general population.

In spite of these findings, other research found a lower incidence of major depression. In a sample of Southwest California Indians and North Plains Indians (the authors used general terms for these tribes as a way to protect the confidentiality of these communities), rates of depression were lower than the general population (Beals et al., 2005). The researchers compared data from the American Indian Service, Utilization, Psychiatric Epidemiology, Risk and Protective Factors Project (AI-SUPERPFP) with data from the baseline National Comorbidity Survey (NCS; Kessler, 1994). The authors hypothesized that they would find comparable or higher rates of major depressive episodes in American Indians than the general population; however, they observed that the lifetime and 12-month major depressive episodes were much lower in the American Indian sample than in the NCS sample. In fact, American Indians' lifetime rates for major depressive episodes were only about 30% of those found in the NCS sample. Based on these findings one would wonder why the authors’ initial assumptions were wrong.

The authors point out cultural factors may play a role in the results of this prevalence study and suggest the operationalization of major depressive episode as a primary factor in nondetection (Beals et al., 2005). The investigators explained that some participants from the North Plains tribe were perhaps less likely to endorse the main major depressive episode symptoms because admitting such symptoms would display emotional weakness. Southwest
California Indians on the other hand, were more likely to endorse depressive episode symptoms, like those in the NCS sample. The data from the AI-SUPERPFP presents one of the most comprehensive assessments of the prevalence of depressive episode in AIs; however, these data were drawn from participants living close to or on the reservations, and may not be generalizable to urban Indians living away from their tribe (Garoutte et al., 2003).

Several studies suggest some American Indian cultures may have a different conceptualization of depression. According to Tremansen and Ryan (1970) “In some Indian languages there is no native term for depression as such, but rather a series of terms applied to various states of sadness, concern, and hopelessness” (p. 125). One study observed considerable differences in the way Hopi Indians perceive depression (Mason, Shore, & Bloom, 1985). The authors suggest that the DSM-III formulation of major depressive disorder does not match up directly to categories of illness for the Hopi.

In view of the fact that some studies (Mason et al., 1985; Chance 1962) have reported differences in the way several American Indian and Alaska Native groups perceive depression, one may wonder about the appropriateness of using certain standardized tests with these groups. However, several studies have used the Center for Epidemiological Studies-Depression Scale (CES-D) with various American Indian populations. These samples include: American Indian adolescents (Dick, Beals, Keane, & Manson, 1994), American Indian college students (Beals, Manson, Keane, & Dick, 1991) American Indian adults (Somervell et al., 1992; Somervell et al., 1993) and American Indian elders (Chapleski et al., 1997).

The conventional approach to administering the CES-D requires users to respond using a paper-and-pencil format. This may not be an ideal approach for all populations, nor is it free from error. In fact, these errors may occur when a client completes an answer sheet such as:
skipping an item, filling in the wrong blank, or failing to mark a response grid fully (Krug, 1987). Considering the potential errors associated with paper-and-pencil test, and the inability of some patients to use them, other approaches need to be considered. One alternative approach to conventional paper-and-pencil test is to use a computerized version of the same test.

Research evaluating the use of computerized testing has found this approach to be a reliable and valid in screening tool for depression in both the clinical and non-clinical settings (Kobak, Reynolds, Rosenfeld, & Greist, 1990; Muñoz et al., 1999). Additionally, investigators have found equivalence of test modalities (paper-and-pencil format and computerized format) for depression screening tools with samples of university students (González, Spiteri, & Knowlton, 1995; Schulenberg & Yutrzenka, 2001). Although these studies have limited generalizability because of the samples, they were stepping stones to the development of later studies using patients from health care settings such as hospitals, mental health centers, and primary care facilities (Muñoz, González, & Starkweather, 1995; Muñoz, McQuaid, González, Dimas, & Rosales, 1999; González et al., 2000; González & Shriver, 2004; González, Carter, & Blanes, 2007). The literature on computerized methods reveals this approach as reliable and valid in a variety of settings.

An important issue to consider when evaluating computerized approaches in research is the participant’s potential preference to using computerized assessments. One study observed 22 of the 27 participants preferring to talk with a practitioner rather than using a computerized format (Greist, Klein, & Van Cura, 1973). However, these results should be interpreted with caution because the study was conducted in a time when computers were not as common or accessible to people as they are today. Interestingly, 21 of the 27 patients thought the computerized format was easy to use. In spite of these early findings, some argue patients may
feel less embarrassed in divulging sensitive information about their mental status using a computerized format (Erdman, Klein, and Greist, 1985; Kobak, Greist, Jefferson, & Katzelnick, 1996).

Another study reported that patients at a high risk of committing suicide with high suicidal ideation appeared to feel more comfortable using a computerized version of the Hamilton Rating Scale for Depression than confiding in a clinician (Levine, Ancil, & Roberts, 1989). Petrie and Abell (1994) observed that 52% of 150 parasuicidal patients (111 female and 39 males; mean age = 29.1) admitted to a New Zealand Hospital preferred to disclose personal information in a computerized interview compared to 17.4% who preferred a clinician. This may suggest certain individuals may prefer a computerized interview due to the objective nature of this format (Butcher, Perry, & Hahn, 2004), whereas with a face-to-face interview they may feel like they are being judged (Erdman et al., 1985).

Research also suggests that individuals may prefer a computerized format when answering questions related to alcohol consumption. In one study, patients revealed they drank alcohol in higher amounts to a computerized-administered interview than to a psychiatrist (Lucas, Mullin, Luna, & McInroy, 1977). Furthermore, these patients displayed a high level of acceptability of the computerized method in a computer attitudes questionnaire.

Participants using a computerized administered version of the Minnesota Multiphasic Personality Inventory (MMPI) displayed less anxiety and felt relaxed using a paper-and-pencil version (Rozensky, Honor, Rasinski, & Tovian, 1986). In the same way, researchers comparing test modalities (computerized vs. paper-and-pencil) of the Eysenck Personality Questionnaire and the Carrol Rating Scale for Depression found that participants accepted the computerized test and the finding was not dependent on past computer experience (Merten & Ruch, 1996).
Clearly, the use of computerized methods can be a method of preference for certain sensitive issues.

Some researchers argue that when evaluating the equivalence of computerized testing to conventional approaches it is important to incorporate a measure for possible apprehensions toward computers by the user as a way to control for this type of confound (Schulenberg & Yutrzenka, 1999; Schulenberg & Yutrzenka, 2001). The most commonly used measures for this construct of computer anxiety are the Computer Anxiety Rating Scale- Form C (CARS-C; Rosen, Sears, & Weil, 1987), the Computer Aversion, Attitudes, and Familiarity Index (CAAFI; Schulenberg & Melton, 2008), and the Computer Aversion Scale (CAVS; Meier, 1988; the CAVS was used for the current study and will be discussed further in the methods section).

Overall, research suggests a computerized depression screening tool is appropriate in a wide variety of settings and can be used with diverse groups of people. However, the evaluation of computerized depression screening with AI/AN populations in a primary care setting is still not well understood. The author is not aware of any studies that have evaluated a computerized approach to screen symptoms of major depression in AI/ANs. However, past studies looking at mental health issues, such as AI-SUPERPFP study and Lapham, Kring, Kleyboeker (1997), have used the approach to facilitate data collection, but were not evaluating the approach.

It is important to evaluate a computerized screening tool with AI/ANs because depression occurs in these communities as they do with other populations. Although past research has revealed mixed findings for depression rates with AI/ANs, other factors closely related to depression, such as substance abuse and suicide are high (Olson & Wahab, 2006). Further, IHS does not provide the funding needed to address the current health care needs of AI/ANs. In fact, it operates at approximately 50% of what is needed for this group (U.S. Commission on Civil
Rights, 2003 as cited in Olson & Wahab, 2006). Nelson et al. (1992) explains that there are unmet mental health staffing needs, where there is only 50% of what is needed to provide adequate services.

Not only is this type of research important for better understanding computerized depression screening, but the data collected can also aid the health institution by providing health related information about the population they serve. In discussing the importance of assessment and intervention research in AI primary care clinics, Gilder et al. (2013) suggest screening can aid, not only to better understand a health problem, but also to help promote education based outreach on depression and the importance of depression screening to the local community. Further, the health institution may benefit from data collected when applying for grant funding to help build infrastructure.

Computerized methods are promising tools addressing a need for innovative approaches for reaching at risk populations. Although much research has been done using a computerized depression screening tool called the Voice Interactive Depression Assessment System (VIDAS) in English- and Spanish-speaking communities, the AI/AN communities have not been examined. The purpose of this study was to evaluate the use of a computerized interactive tool, VIDAS, with individuals from a rural primary care clinic providing services to AI/ANs. To do this, the author examined the psychometric properties of VIDAS, and the participant’s acceptability of computerized testing, the participant’s preference for test modality (computerized vs. paper-and-pencil). It is hypothesized VIDAS would demonstrate:

- high sensitivity (the proportion of correctly classified symptoms to total positive cases) and specificity (the proportion of incorrectly classified symptoms to total negatives cases) for correctly detecting symptoms of depression;
• equivalence to the paper-and-pencil CES-D;
• high inter-item reliability;
• high validity correlations with the paper-and-pencil form BDI-II;
• similar correlations as the paper-and-pencil CES-D format with the Computer Aversion Scale (CAVS) total score;
• more favorable participant responses over the paper-and-pencil approach with the Preference for Mode of Administration Questionnaire (PMAQ).

Methods

Participants

One hundred participants (30 males and 70 females) were obtained from a rural clinic in California. Two female participants’ interview data were removed because they displayed symptoms of schizophrenia - delusional statements and trouble staying on task. All who participated identified as American Indian or Alaska Native (AI/AN) and were between the ages of 18 to 77 ($M = 39.40$, $SD = 13.00$) years old, with 52% living on a reservation. The level of education achieved varied among participants, with most reporting some college experience/associate degree (46%), followed by: high school diploma/ G.E.D (26%), less than high school (15%), college graduate (12%), other (1%). The level of computer experience also varied among participants, with most reporting some level of experience (48%), followed by: experienced (34%), very experienced (10%), and none (8%).

Materials

The following materials were used: an informed consent, a demographic questionnaire, a laptop computer, the Voice Interactive Depression Assessment System (VIDAS), the Center for Epidemiologic Studies-Depression scale (CES-D), the Beck Depression Inventory (BDI-II), the
Patient Health Questionnaire (PHQ-9), the Computer Aversion Scale (CAVS), and the Preference for Mode of Administration Questionnaire (PMAQ).

**Demographic questionnaire.** The Demographic Questionnaire included items addressing the participants’ age, gender, and current residence, level of education, tribal background, and computer experience.

The Voice Interactive Depression Assessment System (VIDAS). This study used VIDAS for the computerized test condition. VIDAS uses the 20-item questions from the CES-D in a voice interactive format, where participants’ verbal responses are recorded on a laptop computer using a telephone receiver. VIDAS employs a digital image of an interviewer who verbally presents CES-D questions. The digitized interviewer was randomized in either a male or a female format (See Figure 1 for screenshot of digitized male interviewer). Thus, all participants had an equal chance, regardless of their gender, of seeing a male or female digitized interviewer. The male and female characters presented in digitized form were created using male and female Hispanic actors. The gender of the digitized interviewers revealed no significant difference in the comfort ratings of participants (González et al., 2007).

In terms of its use in past research, González and Shriver (2004) determined that internal consistency of VIDAS subscales (CES-D and DSM-IV) were good to excellent in both English and Spanish (0.81 to 0.92) with adequate criterion validity (0.58 to 0.67). In addition, the investigators found VIDAS subscales displayed sensitivity (0.72 to 1), and specificity (0.40 to 0.72) to accurately detect symptoms of depression.

In a more recent study, González et al. (2007) also determined good internal consistency for VIDAS ranging from good to excellent in both English and Spanish (0.68 to 0.90). Convergent validity of VIDAS total scores with the BDI-II total scores was moderate to strong.
for all groups (0.41 to 0.80). Overall, VIDAS has previously demonstrated adequate to excellent test performance with diverse samples.

The Center for Epidemiologic Studies-Depression scale (CES-D). The CES-D is a 20-item survey developed by the Center for Epidemiologic Studies for screening depressive symptoms in the last week and was designed to be used with community populations (Radloff, 1977; Weissman, Sholomskas, Pottenger, Prusoff, & Locke, 1977; Roberts, 1980), and has also been tested in clinical populations (Radloff, 1977; Weissman, et al., 1977; Roberts, Vernon, & Rhoades, 1989). Each questionnaire item has a possible value ranging from 0 - 3 (0 = rarely or none of the time, 1 = some of the time, 2 = a lot of the time, and 3 = most of the time), and scores can range from 0 to 60. The positively worded items are reverse coded, which include items 4, 8, 12, and 16. A cutoff score of 16 is often used as a way to indicate depressive symptoms (Weissman et al., 1977), and the items in the measure fall into four factors (i.e., negative affect, positive affect, somatic complaints and interpersonal relations; Radloff, 1977).

Early research found that the CES-D demonstrated high internal consistency for community and clinical populations (range 0.83-0.91), adequate test-retest reliability (0.32-0.70), and moderate to high correlations with other depression measures (Radloff, 1977; Weissman et al., 1977). The CES-D has been tested extensively with diverse populations including: a community sample of African American, Mexican Americans, and Anglo-American adults (Roberts, 1980); and Mexican (English and Spanish-speaking) or Anglo-American psychiatric inpatients (Roberts et al., 1989); Anglo-American, U.S. born Mexican, Mexico born Mexican (Golding & Anhensel, 1989); urban Latinos (Posner, Stewart, Marin, & Perez-Stable (2001); married Chinese couples (Cheung & Babley, 1998); American Koreans and Anglo Americans (Kim, Han, & Phillips, 2003); Korean immigrants (Kim, Seo, & Cain, 2010); and a community
sample of Vietnamese adults (Tran, Ngo, & Conway, 2003). As mentioned earlier, the CES-D has also been tested with American Indian adolescents (Dick, Beals, Keane, & Manson, 1994), American Indian college students (Beals, Manson, Keane, & Dick, 1991) American Indian adults (Somervell et al., 1992; Somervell et al., 1993) and American Indian elders (Chapleski et al., 1997).

The Beck Depression Inventory-II (BDI-II). The BDI-II is a highly reliable multiple choice 21-item questionnaire measuring both the characteristic attitudes and symptoms of depressed mood (Beck, Steer, & Brown, 1996), and is considered the preeminent standard for depression screeners. Each question has a set of 4 possible answer choices ranging from 0 to 3, giving a total score of possible depression symptom levels. The cutoffs for depression symptom scores are: 0–13: minimal depression; 14–19: mild depression; 20–28: moderate depression; and 29–63 severe depression. Higher total scores indicate more severe depressive symptoms.

The BDI-II has demonstrated strong psychometric properties with various samples, including college students, adult psychiatric outpatients (Beck et al, 1996); youth nonclinical (Osman, Barrio, Gutierrez, Williams, & Bailey, 2008), psychiatric outpatients (Steer, Kumar, Ranieri, & Beck, 1998) and psychiatric inpatients (Krefetz, Steer, Gulab, & Beck, 2002; Kumar, Steer, Teitelman, & Villacis, 2002); primary care patients (Arnau, Meagher, Norris, & Bramson; 2001); and with diverse populations (Joe, Woolley, Brown, Gharamanlou-Holloway, & Beck, 2008; Al-Musawi, 2001; Byrne, Steward, & Lee, 2004). Internal consistencies from the aforementioned studies have ranged from 0.84 to 0.94, and the instrument has demonstrated convergent validity with various depression, anxiety, and suicide measures.

The Patient Health Questionnaire (PHQ-9). The PHQ-9 is a self-report measure specific to depression derived from the Patient Health Questionnaire (PHQ; Spitzer et al., 1999). This
instrument scores the 9 DSM-IV criteria based on the mood module from the original PRIME-MD (Spitzer, Kroenke, & Williams, 1999), and can be used as a diagnostic tool. In order to reach a diagnosis of major depression, one must have 5 or more depressive symptom criteria existing for the minimum of “more than half the days” in the previous 2 weeks, and one of the symptoms must be anhedonia or depressed mood (Kroenke, Spitzer, Janet, & Williams, 2001). As a diagnostic tool, the PRIME-MD has sensitivity and specificity equivalent to the diagnosis made by a mental health professional (Spitzer et al, 1994). Kroenke et al. (2001) found the area under the curve (AUC) for the PHQ-9 in diagnosing depression was 0.95. Overall, research has found the PHQ-9 to be a valid and reliable measure in making a diagnosis and in assessing depression severity (Kroenke et al., 2001). The PHQ-9 may also provide a measure of depression severity, where 1 to 4 indicates no depression, 5 to 9 indicates mild depression, 10 – 14 indicates moderate depression, 15 – 19 reveals moderately severe depression, and 20 to 27 indicates severe depression (Kroenke & Spitzer, 2002). It should be noted the severity scores may vary among participants where two individuals may have the same score near the cutoff for diagnosis but one may not be diagnosed because they did not answer “more than half the days” in previous 2 weeks for either the anhedonia or depressed mood questions.

The PRIME-MD has been tested with a sample from a Indian Health Service primary care clinic in Albuquerque, NM (Parker et al., 1997). Parker et al. (1997) observed that out of 100 patients evaluated using the PRIME-MD, 10 (10%) individuals were identified with major depressive disorders. These results were similar to Spitzer’s (1994) findings in the PRIME-MD 1000 study where 115 (12%) individuals in sampled had major depressive disorders. However, while Parker et al. (1997) determined fair agreement between the diagnoses from the PRIME-MD with a mental health professional with a kappa coefficient of 0.56 (accuracy rate = 79%),
Spitzer (1994) observed better agreement with a kappa coefficient of 0.66 (accuracy rate = 91%). Overall, the results suggest the PRIME-MD maybe an appropriate diagnostic tool with American Indian communities.

_The Computer Aversion Scale (CAVS)._ The CAVS is a 31-item true-false measure designed to examine one’s willingness to use computerized methods. Those who score above a 19 are believed to have an aversion towards using computers. Meier and Lambert (1991) suggest the CAVS can help measure the construct of anxiety towards using computers. Each item is in the form of a “true” or “false” response, and the measure is scored with a total score and three subscale scores. Past research has revealed high alpha coefficients of 0.89 for the total score, 0.80 for the Efficacy score, 0.81 for the Outcome score, and 0.74 for the Reinforcement score (Meir, 1988). Meier and Lambert (1991) later found for 1,234 participants test-retest reliabilities for the CAVS ranging from 0.74 to 0.78 across three time points. The authors also found that for each time point the CAVS and CARS were correlated, with a range of 0.62 to 0.67. The researchers determined the internal consistency for a new subscale called the Negative Feelings for Computer ranged from 0.68 to 0.87 (Meier, 1988). Schulenberg and Yutrzenka (2001) also found support for the reliability of CAVS with alphas of 0.87, 0.74, 0.68, 0.77 for CAVS total scores, Outcome scores, Reinforcement Scores, and Efficacy Scores, respectively.

_The Preference for Mode of Administration Questionnaire (PMAQ)._ The 8-item PMAQ was created by Merten in 1994 and published by Merten and Ruch (1996; the original version had 9-items but the authors suggested removing question 5 to increase the psychometric properties of the measure). Each question is scored as a 1 (_preference for conventional assessment_), a 2 (_no preference_), or a 3 (_preference for computerized assessment_). The PMAQ is understood by looking at whether an individual’s obtained score is significantly different from
the expected score of 16. A total score of 16 would indicate indifference toward test format, while below 16 would indicate a preference for paper-and-pencil formats and an above 16 would indicate a preference for computerized formats. Merten and Ruch (1994) found an internal consistency of 0.83 for the PMAQ.

**Procedure**

Two Institutional Review Boards (IRBs) reviewed and approved applications for the implementation of this study. These IRBs were the California State University, San Marcos IRB and the Southern California Tribal Health Center (SCTHC) IRB. All research activities and protocols were approved by the IRBs as a way to ensure participants’ rights were preserved. All study interviewers, including two research assistants, completed the Collaborative Institutional Training Initiative’s Human Subjects Research Curriculum and were trained in how to collect data and follow the study protocol in a culturally appropriate manner.

SCTHC offers a wide variety of services including podiatry, ophthalmology, and clinics in obstetrics and gynecology and diabetes as well as community health service. Participants were recruited in the Behavioral Health, Medical, and Dental departments from the SCTHC located in California. Criteria for study participation included that the participants needed to be 18 years of age or older, be an enrolled tribal member or a descendent of an enrolled tribal member from a Federally recognized tribe within the United States.

After the research participant agreed to participate, the interview was conducted in a private room located in the clinic. The research staff explained the informed consent and the rights study participants have in terms of being able to skip questions they do not want to answer, and indicated they could withdraw at any time during the study without penalty and still receive a gift card. Participants then completed a demographic questionnaire in an interview format.
Following the demographic questionnaire, participants completed four depression measures in random order to control for order effects. Following these tests, participants completed the remaining measures. The interviews lasted approximately 22-40 minutes.

As part of the research protocol, participants who reported suicidality completed a risk assessment to determine if there was a significant threat. Those who were suicidal or had high levels of depression were offered help in setting up an appointment with mental health staff in the SCTHC Behavioral Health Department. Following the testing, participants were debriefed concerning their participation in the study. They were given a $10 Target gift card and were offered a list of Behavioral Health services at SCTHC.

Data collection lasted approximately 7.5 months from August 2012 to April 2013. On average, five to 14 interviews were conducted per week, but varied during the holidays, the time of day for recruitment, and the availability of research staff to conduct interviews.

Sample Size and Power

Initially, the author proposed to collect a sample size of 130 participants, but could not reach this amount. The original proposed sample was based on a power analysis for a mixed study design. This mixed design included between subjects (depressed vs. nondepressed as determined using the PHQ-9) and within-subjects components. For a medium effect size for the mixed design, the author would have needed 65 depressed and 65 non-depressed participants at Power = .80 for $\alpha = 0.05$ (Cohen, 1991). However, the author encountered significant recruitment challenges reaching 130 interviews, and in obtaining 65 depressed individuals.

To recruit 130 individuals overall with 65 depressed cases was challenging because not many individuals were willing to participate. In talking with several community members and other researchers conducting studies at SCTHC, there appears to be a certain level of distrust
when it comes to research. This distrust appears to be related to: fears of a loss of confidentially, where personal information that the participant discloses would be revealed to others in the community; researchers may want to conduct research, but not give anything back to the community after it is over; and the research conducted could portray the community as a whole in an unfavorable way. After review of this issue, the thesis committee agreed with reducing the sample size to 100 participants. Given there was reduction in sample size and the groups were not equal in size, there is a reduction of the power of the study and thus the results for the ROC analysis should be interpreted with caution. However, all other analyses have sufficient power because the sample size is appropriate for them.

**Results**

This section contains information on the data collected and the analyses used to understand the accuracy of the depression measures, equivalence of test formats (computerized and paper-and-pencil), the internal consistency and convergent validity of the depression measures, the acceptability of computerized testing to participants, and participant test format preference. Further, the author will provide background information on the ROC analysis and Bayes factor $t$ test.

The accuracy of computerized screening test correctly recording the participants’ responses was 89% with VIDAS scores ranging from 1 to 44 ($M = 16.42$, $SD = 11.02$). The other measures are as follows: paper-and-pencil CES-D scores ranged from 0 to 44 ($M = 15.43$, $SD = 11.90$), BDI-II scores ranged from 0 to 51 ($M = 12.48$, $SD = 11.34$), and PHQ-9 scores ranged from 0 to 24 ($M = 6.10$, $SD = 5.96$). The frequencies of depression symptoms of a score of 16 and above for VIDAS was 47% and was 37% for the CES-D. Seventeen participants scored in the major depression category using the DSM-IV criteria on the PHQ-9. Additionally,
depression severity was observed by looking at the participant’s overall PHQ-9 score. See Table 1 to see the frequency of severity of PHQ-9 scores by gender.

**Psychometric Properties of Depression Screeners**

A nonparametric receiver operating characteristic (ROC) analysis was used to interpret the accuracy of the depression screening tools (VIDAS, CES-D, BDI-II) by examining their area under the curve (AUC) values, sensitivity and specificity, cut scores, and predictive values. For this study, ROC analyses were conducted using the VIDAS total scores, CES-D total scores, and BDI-II total scores with the PHQ-9 major depression cases. The PHQ-9 was used as a “gold standard” which the screeners were tested against to run the ROC analyses.

The x-axis of an ROC curve shows the false-positive error rate (1-specificity) and the y-axis shows the sensitivity of a test. Sensitivity reveals the likelihood of the measures accurately recognizing depressive symptom levels when a participant actually has major depression (also expressed as positives correctly classified / total positives; Fawcett, 2006). Specificity reveals the likelihood of the measures accurately recognizing the absence of depressive symptoms when the participant does not have major depression (negatives incorrectly classified / total negatives; Fawcett, 2006). An ROC curve can reveal a good cutoff point for sensitivity and false-positive error rate.

The area under the ROC curve is a single value reflecting a measure’s performance. It also explains the compromise between sensitivity and false positive rate (1-specificity; Metz, 1978), and is similar to the Wilcoxon test of ranks (Hanley and McNeil, 1982). The closer an AUC value is to 1, the more accurate the screening measure is. No accurate measure would have a value less than 0.50 (Obuchowski, 1997).
Many studies evaluating depression screening tools with ROC analysis also look at predictive values. Predictive values give a probability of a correct recognition by the measure, which sensitivity and specificity do not (Altman & Bland, 1994). A positive predictive value (PPV) is the proportion of observations with a positive test results who are correctly recognized \[
\text{PPV} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}},
\]
whereas negative predictive value (NPV) is the proportion of observations with negative test results who are correctly recognized \[
\text{NPV} = \frac{\text{true negative}}{\text{false negative} + \text{true negative}}.
\]

The computer program SPSS version 17.0 was used to conduct the ROC analyses. The AUC value for VIDAS was 0.97 (95% CI 0.94 to 1). The AUC values for the CES-D and the BDI-II were 0.97 (95% CI 0.93 to 1) and 0.95 (95% CI .91 to 1), respectively. Looking at the ROC curve revealed that an optimal cutoff score of ≥ 22 had the ideal sensitivity of 0.94 and specificity of 0.89 (See Figure 2 for ROC curves for VIDAS). The predictive performance of VIDAS for a cutoff of ≥ 22 had a PPV of 64% and a NPV of 99%, while a cutoff of 16 had a 37% positive predictive value (PPV) and 100% negative predictive value (NPV). An ROC curve for paper-and-pencil CES-D revealed an optimal cutoff score of ≥ 18, with sensitivity of 0.94 and specificity of 0.81, and an optimal cutoff score of ≥ 15 for the BDI-II, with sensitivity of 0.94 and specificity of 0.85 (See Figure 3 for an ROC curve for the CES-D and Figure 4 for an ROC curve for the BDI-II). The predictive performance of the CES-D for a cutoff of ≥ 18 had a 52% positive predictive value (PPV) and 99% negative predictive value (NPV). The predictive performance of BDI-II for a cutoff of ≥ 15 had a PPV of 57% and a NPV of 99%.
Equivalence of Test Formats (VIDAS and CES-D)

Studies examining the equivalence of computerized and paper-and-pencil formats have typically used traditional hypothesis testing procedures [Null Hypothesis Significance Testing (NHST)] to test for significance (i.e., using a t-test to determine whether the means of two groups of scores are significantly different). In studies of equivalence, the desired outcome is to find no differences between formats and, therefore, statistically the hope is not to find a significant difference. Concluding there is no differences in means as a way to demonstrate equivalence of formats (computerized and paper-and-pencil) is a way of saying the data support the null hypothesis. Critics have argued, however, that NHST is not the appropriate statistical methodology when researchers hope to provide support for the null hypothesis (Rouder, Speckman, Sun, Morey, and Iverson, 2009). In fact, one cannot support the null using NHST, only fail to reject it.

What equivalence studies require is a way of providing the probability the null hypothesis (no difference) is true, given a particular data set \(p(H|D)\). While NHST is not appropriate for this purpose, Bayesian analysis is a suitable approach (Gallistel, 2009; Rouder et al. (2009). One statistic coming of the Bayesian approach is the Bayes factor, which is the ratio of the likelihoods of the two theories under consideration.

Using a Bayes factor allows the data to support either the null or alternative hypothesis. Rouder et al. (2009) developed a Web-based-program (available at pcl.missouri.edu) to convert information from a within subjects t-test into a Bayes factor (Rouder et al., 2009). The online program minimizes the difficulty involved in the computation of a Bayes factor and allows the user to input values for \(t\) (conventional t statistic), \(N\) (the number of observations or sample size), and a \(r\) value for effect size (which is set to 1.0 and is the recommended value by the authors as
an objective prior), which in turn will produce what is known as a JZS Bayes factor. This statistic uses the JZS prior, which provides the parameter of variance for the null hypothesis and the parameter of variance and effect size for the alternative hypothesis (Morey & Rouder, 2011). Rouder et al. (2009) also provides a table of critical $t$ values to interpret the JZS Bayes factor values, which can explain what the data favors, the null or alternative hypothesis.

For the current study, the null hypothesis was there would be no differences in means for both test modalities (VIDAS and paper-and-pencil CES-D), while the alternative hypothesis was there are differences in means of the test formats. The assumption of mean differences for VIDAS and the paper-and-pencil CES-D revealed a violation of normality using a Shapiro-Wilk test, $p = 0.038$.

The information from a dependent-samples $t$ test was used for a JZS Bayes factor. The dependent-samples $t$ test revealed no significant mean differences for test modalities (VIDAS and paper-and-pencil CES-D), with $t (97) = 1.64$, $p > 0.05$, $d = 0.17$ (95% CI 0.21 to 2.19). Looking at a $t$ value tables by Rouder et al. (2009, p. 232), a JZS Bayes factor (using a an objective Bayes approach with a JZS prior) of 3.37 revealed favor for the null hypotheses of no differences, meaning the data are approximately 3 times more likely under the null hypothesis than the alternative hypothesis. These results suggest that the test formats are equivalent.

*Reliability of Depression Measures*

Cronbach’s alphas for the depression measures are as follows: 0.88 for VIDAS items, 0.92 for pencil-and-paper CES-D items, 0.93 for the paper-and-pencil BDI-II items, and 0.88 for face-to-face interview PHQ-9 items.
Validity of VIDAS

Convergent validity was examined by looking at correlations between VIDAS total scores with face-to-face PHQ-9 total scores, VIDAS total scores with paper-and-pencil CES-D total scores, and VIDAS total scores with paper-and-pencil BDI-II total scores. First, the correlation coefficient between VIDAS total scores and PHQ-9 total scores was strong and positive (\( r = 0.81, N = 98, p < 0.01, \text{two-tailed} \)) with a correlation of determination (\( r^2 = 0.66 \)) revealing 66% shared variance for VIDAS and PHQ-9 scores. Secondly, the correlation coefficient for the VIDAS total scores and paper-and-pencil CES-D total scores was strong and positive (\( r = 0.87, N = 98, p < 0.01, \text{one-tailed} \)) with a correlation of determination (\( r^2 = 0.75 \)) revealing 75% shared variance for VIDAS and CES-D scores. Lastly, the correlation coefficient for VIDAS total scores and BDI-II total scores was strong and positive (\( r = 0.77, N = 98, p < 0.01, \text{one-tailed} \)) with a correlation of determination (\( r^2 = 0.59 \)) revealing 59% shared variance for VIDAS BDI-II scores.

Results of the Computer Aversion Scale (CAVS) and Preference for Mode of Administration Questionnaire (PMAQ)

The CAVS scores overall ranged from 1 to 24 (\( M = 8.04, SD = 5.50 \)). The CAVS subscales were as follows: Outcome \( M = 3.30, SD = 2.50 \); Efficacy \( M = 2.01, SD = 2.21 \); Reinforcement \( M = 2.72, SD = 1.83 \). The internal consistency of all CAVS items were 0.88, with subscales of 0.56, 0.81, 0.82, 0.75 for Reinforcement scores, Outcome scores, Efficacy scores, Negative Feelings scores, respectively. In addition, age was positively associated with CAVS total scores (\( r = 0.29, N = 98, p = .004, \text{two-tailed} \)), so as age increased so did CAVS total scores with a correlation of determination (\( r^2 = 0.08 \)) revealing little (8%) shared variance for participants’ age with CAVS scores.
CAVS Correlational Analyses with Depression Measures

Correlational analyses were used to determine the similarity of correlations for depression measures total scores with CAVS total scores. There were similar weak positive correlation coefficients for both test modalities with CAVS, VIDAS total scores and CAVS total scores ($r = 0.10, N = 98, p > 0.01$) and paper-and-pencil CES-D and CAVS total scores ($r = 0.14, N = 98, p > 0.01$). Thus, both modalities had similar correlations with CAVS scores. In addition, there were no significant differences for these two correlations ($z = -0.28, p > 0.01$).

The PMAQ scores ($N = 96$) ranged from 8 to 24 ($M = 16.33, SD = 4.49$), and had an internal consistency of 0.85. The correlation of (0.08) between age and PMAQ scores was not significant. See Table 2 for the frequency of test modality preference scores by age. For male participants 13% preferred paper-and-pencil testing, 12% preferred computerized testing, and 5% were indifferent to modality. For female participants 33% preferred paper-and-pencil testing, 33% preferred computerized testing, and 4% were indifferent to modality. Overall, 46% preferred paper-and-pencil testing, 45% preferred computerized, and 9% were indifferent to modality.

Discussion

The goal of this study was to evaluate VIDAS as an alternative screening approach to conventional paper-and-pencil testing with a sample from a primary care clinic servicing AI/AN. To do this, the author examined the psychometric properties (predictive performance, equivalence of formats, internal consistency, and convergent validity), a participant’s acceptability of computerized testing, and his or her preference for test modality (computerized vs. paper-and-pencil).
An ROC analysis revealed that VIDAS, CES-D, and BDI-II all had very good accuracy
in for detecting depression symptoms. Specifically, the finding supports the first hypothesis of
high sensitivity and specificity for VIDAS correctly detecting symptoms of depression.
Additionally, the ROC curve reveals ≥ 22 score is the optimal cutoff. Both findings are
consistent with past research regarding the accuracy of VIDAS in screening participant
symptoms (González & Shriver, 2004; González et al., 2007). The PPVs and NPVs for all the
depression measures using the typical cutoff scores are similar to past research for those
measures; however, this current study found a cutoff score of ≥ 22 yields higher predictive
values than the cutoff of 16. The CES-D also had good accuracy, but had a lower optimal cutoff
score ≥ 18. The differences in optimal cutoff score for the measures are interesting as both tests
modalities contain the same items. One reason for this disparity of cutoff scores differences may
be due to the accuracy of VIDAS in correctly identifying responses, which was 89%. Although
this is excellent accuracy, 100% may have led to closer optimal cutoff scores for the two
measures. Potential problems encountered with speech recognition technology may have to do
with the way it records participant responses. González and Shriver (2004) explain that potential
problems that may affect the way this technology records participant responses may include:
background noise, the pronunciations of the interviewee, and the computer’s performance.

Hypothesis 2 was supported as the JZS factor revealed equivalence of test modalities
(VIDAS and the paper-and-pencil CES-D). Specifically, JZS Bayes factor favored the null of no
differences between testing formats. These findings are important because other research in the
area has typically used NHST, which critics have argued is not the appropriate statistical
methodology when researchers hope to provide support for the null hypothesis (Rouder, et al.,
2009). The current study maybe the first to use a Bayesian approach to examine the equivalence
in depression screening formats. Further equivalence studies should consider using a Bayesian approach that allows support for the null.

In terms of the reliability and validity of VIDAS, the author found strong internal consistency and convergent validity supporting hypotheses 3 and 4. These findings are similar to past studies observations of adequate to strong internal consistency and convergent validity for VIDAS (González and Shriver, 2004, González et al., 2007) and for similar computerized voice recognition CES-D based instruments (González et al., 1995; Munoz et al., 1995; Munoz et al., 1999). Furthermore, the results of VIDAS and the CES-D are similar to and in some cases better than past studies psychometric properties of the CES-D.

Although it is important that a computerized test such as VIDAS display excellent psychometric properties to show the validity of the measure, it is also important to consider the participant responses to using a computerized test. Both the CAVS and PMAQ are important measures past studies have found to address one’s acceptability and possible preference for computerized testing.

The author found support of hypothesis 5 for which total scores from both test modalities (VIDAS and CES-D) displayed similar correlations with the total scores from the CAVS measure. In addition, the means and standard deviations of CAVS total scores for this current study were similar to Schulenberg and Yutrzenkas’ (2001) study; however, they were different from Meier’s (1988) study. The internal consistency of CAVS for this study was consistent with previous studies (Meier, 1988; Schulenberg & Yutrzenka, 2001) for CAVS total scores and Subscales.

As for participant preferences for test modality, Hypothesis 6 was not supported, as participants did not prefer a computerized test. Instead, the frequency of differences was
approximately even for paper-and-pencil and computerized testing preference, with an equal number of participants preferring both modalities. These findings are different from past research which reported a preference by participants for computer administration over paper-and-pencil administration (Schulenberg & Yutrzenka, 2001).

In terms of the overall internal validity of the study, the depression measures all performed well and the computerized measure was equivalent to paper-and-pencil testing. In addition, the author reduced measurement error by a) using multiple measures of the same construct of depression, b) properly training interviewers, c) and double-checking the data thoroughly. It can also be argued the current study had good external validity as the participants were recruited from the SCTHC primary care clinic, and may provide a stepping-stone to future research involving more advanced technology. The accuracy of VIDAS correctly recording the participants’ responses was very good (89%), however, new technology may move accuracy closer to 100%.

Comorbidity

Although the focus of this study was to evaluate a computerized interactive depression screening tool, comorbidity issues (anxiety; Sanderson, Beck, & Beck, 1990; substance dependence; Chen et al., 2013) exist and should be addressed as they can affect one’s score on a depression measure. Thus, comorbidity issues with relation to anxiety or substance use can be confounds to a study and should be controlled. Future studies should attempt to control for these potential confounds methodologically by including additional measures.

Limitations

A limitation of the current study was the sample size of 130 participants was not met, and there were not equal sized comparison groups for testing the accuracy of VIDAS. Thus, the
power of the study for the ROC analysis was reduced and the findings for this part of the study
should be viewed with caution. Nonetheless, despite a reduction in power, the AUC values were
very high and should be examined further for future research.

This study appears to show depression symptoms were stable across the various measures
in the study and did appear to actually measure depression symptoms. However, the PHQ-9
revealed 17 participants had major depression, yet when looking at depression severity for those
cases only 12 individuals had moderately severe to severe levels of depression, while 4
individuals had a moderate level and 1 had a mild level of depression severity. It is likely
participants who scored in the moderate range may be at the higher end of this level. This
discrepancy can cause confusion for the investigator. Although it is efficient to use the PHQ-9, a
more comprehensive measure such as the PRIME-MD, which is helpful for primary care
research, or the Structured Clinical Interview for DSM Disorders I (SCID I; First, Spitzer,
Gibbon, & Williams, 1997), which has versions for primary care research and psychiatric
patients, may yield more accurate diagnosis. However, the SCID versions can be time intensive
and cost-prohibitive for preliminary research.

Another possible limitation had to do with the design of the study. Although the author
randomized the depression measures to control for order effects, future studies should consider
having participants take the test on separate occasions to increase the strength of the study.
Moreover, future studies should consider comparing samples of different clinics serving AI/ANs
from various geographic areas. It would be beneficial to compare Indian Health Service clinics
in regions in urban and rural areas.
Conclusion

In conclusion, evaluation of VIDAS with an AI/AN sample from an IHS clinic used an appropriate sample size for all the analyses except for the ROC. Although the original intent was to obtain a sample size of 130 participants, the results still reveal VIDAS was equivalent to paper-and-pencil version of the CES-D for the within subjects component, and was reliable and valid. The findings also suggest no significant levels of apprehension toward using computerized testing, and participants were equally likely to prefer either test modality.

The implications for the findings of this study suggest that a computerized screening tool can offer an alternative approach to conventional testing in detecting depression symptoms in a patient population. A computerized screening test model such as VIDAS can be easily integrated in a clinical setting by the use of an electronic device such as a laptop computer or tablet computer. Technology is advancing very rapidly, and primary care institutions could easily integrate a computerized approach with electronic patient records. Additionally, practitioners would track a patient’s mental health over time to increase positive health outcomes by using a quick computerized screening tool like VIDAS. Overall, as technology progresses in depression assessment and screening, more studies will be needed to evaluate their psychometric properties and ease of use by potential users of the measures.
References


doi:10.1177/073999863950172004


doi:10.1177/014662167700100306


Table 1

*Depression Severity by Gender*

<table>
<thead>
<tr>
<th>Level of Depression Severity, PHQ-9 scores (N = 98)</th>
<th>Gender</th>
<th>Minimal, 0-4</th>
<th>Mild, 5-9</th>
<th>Moderate, 10-14</th>
<th>Moderately Severe 15 – 19</th>
<th>Severe 20-27</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>19 (19%)</td>
<td>6 (6%)</td>
<td>2 (2%)</td>
<td>1 (1%)</td>
<td>2 (2%)</td>
<td>30 (31%)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>32 (33%)</td>
<td>19 (19%)</td>
<td>8 (8%)</td>
<td>7 (7%)</td>
<td>2 (2%)</td>
<td>68 (69%)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>51 (52%)</td>
<td>25 (26%)</td>
<td>10 (10%)</td>
<td>8 (8%)</td>
<td>4 (4%)</td>
<td>98 (100%)</td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Depression severity was determined by an interview using the Patient Health Questionnaire-9 (PHQ-9; Kroenke, Spitzer, Janet, & Williams, 2001) measure.
Table 2

Preference of Test Modality by Age

<table>
<thead>
<tr>
<th>PMAQ</th>
<th>18-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55+</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper-and-pencil</td>
<td>10 (10%)</td>
<td>10 (10%)</td>
<td>10 (10%)</td>
<td>9 (9%)</td>
<td>6 (6%)</td>
<td>45 (46%)</td>
</tr>
<tr>
<td>Computer</td>
<td>4 (4%)</td>
<td>12 (12%)</td>
<td>9 (9%)</td>
<td>14 (14%)</td>
<td>5 (5%)</td>
<td>44 (45%)</td>
</tr>
<tr>
<td>No Difference</td>
<td>1 (1%)</td>
<td>1 (1%)</td>
<td>6 (6%)</td>
<td>1 (1%)</td>
<td>0 (0%)</td>
<td>9 (9%)</td>
</tr>
</tbody>
</table>

Note: Preference of test modality was determined by the Preference for Mode of Administration Questionnaire (PMAQ; Merten & Ruch, 1996).
Figure 1. Screenshot of the Voice Interactive Depression Assessment System (VIDAS) Interface with Digitized Male Interviewer.
Figure 2. Receiver Operating Characteristic (ROC) for the Voice Interactive Depression Assessment System (VIDAS).
Figure 3. Receiver Operating Characteristic for the Center of Epidemiological Studies-Depression (CES-D) scale.
Figure 4. Receiver Operating Characteristic for the Beck Depression Inventory-II (BDI-II).