Behavioral Screening Measures Delivered With a Smartphone App

Psychometric Properties and User Preference

Nigel E. Bush, PhD, Nancy Skopp, PhD, Derek Smolenski, PhD, Rosa Crampton, MBA, RN, and Jonathan Fairall, BS

Abstract: The smartphone is an increasingly widespread technological vehicle for general health and psychological health promotion, evaluation, education, and sometimes intervention. However, the psychometric performance of behavioral health screening measures has not been commonly evaluated for the new, small-format, touch-screen medium. Before mobile-based applications for behavioral health screening can be disseminated confidently, the reliability and the validity of measures administered by the smartphone must be evaluated. We compared psychometric properties (i.e., internal consistency and test-retest reliability) of seven behavioral health measures completed on paper, a computer, and an iPhone by 45 army soldiers. The results showed the internal consistencies of the smartphone-delivered measures to be equivalent and very high across all three modalities and the test-retest reliability of the iPhone measures also to be very high. Furthermore, completion of the behavioral screening measures by the iPhone was highly preferred over the other modalities and was reported to be easy and convenient. Our findings help corroborate the use of smartphones and other small mobile devices for behavioral health screening.

Key Words: Military, behavioral health, smartphones, psychometrics.

(J Nerv Ment Dis 2013;201: 991–995)

Psychological ailments, behavioral disorders, and traumatic brain injuries (TBIs) among military service members and veterans have increased substantially since the start of the wars in Iraq and Afghanistan (Seal et al., 2009; Terrio et al., 2009; Thomas et al., 2010). Behavioral health problems among active service members and veterans have encompassed anxiety (Stecker et al., 2010) and depression (Tanielian and Jaycox, 2008); stress-related sleep issues (Peterson et al., 2008); substance abuse (Dervaux and Laqueille, 2008; Fear and Wessely, 2009); unprecedented levels of suicide (Kuehn, 2010); and, notably, posttraumatic stress disorder (PTSD; Milliken et al., 2007; Thomas et al., 2010; Vasterling et al., 2010).

Behavioral health specialists typically employ widely used and well-established standardized screening measures in their evaluations of military service members and veterans. For example, the military services use electronic screening packages of self-assessment questionnaires delivered by a computer or a computer kiosk after a deployment (Department of Defense [DoD] Post-Deployment Health Assessment [PDHRA] Program; DoD, 2012). User responses to electronic versions of standardized measures of behavioral functioning are scored, interpreted, and made available to providers.

Although comprehensive computerized systems work well when access is available to large medical treatment facilities, a much more convenient and readily accessible resource would benefit service members on deployment and in other less well-served situations. To that end, the Defense Department’s National Center for Telehealth and Technology (T2) (The National Center for Telehealth and Technology, 2012) is developing a “Mobile Screener” smartphone tool for deployment. The Mobile Screener will deliver a brief core of key behavioral health screening measures to a highly mobile military population in support of a number of scenarios. When used in deployed or field environments where provider access is available only remotely, the Mobile Screener can help gather standardized information about an individual’s behavioral health “vital signs” for referral to indicated care. The Mobile Screener can also be incorporated into in-person care as a treatment-outcome tracking tool that the service member completes between sessions or for use in clinics/waiting rooms as a more portable tool for automated symptom assessment before an appointment.

For the initial phase of the Mobile Screener development, we convened a panel of subject matter experts to determine the core content of the application. In an iterative process over several months, the panel conducted extensive literature reviews, consulted key informants, and examined psychometric use data from self-administered military screening instruments and other resources. The panel selected seven measures to encompass as broad a range as possible of military service–relevant critical behavioral health indices while keeping the content brief. The core measures assessed PTSD, depression, suicidal ideation, unit support, sleep, anger, and TBI.

Although the reliability of the seven measures we chose for the Mobile Screener has been established when administered by conventional modalities, to our knowledge, these have not been evaluated for delivery by smartphone. The following brief report describes the internal consistency, test-retest reliability, usability, preferences, and user experience of the seven core measures of the Mobile Screener delivered by iPhone and the comparative reliability among the iPhone, computer, and paper modes of administration.

METHODS

Design

We used a cross-sectional design that compared the performance of the core behavioral health measures completed on three different modalities in randomized order: a) by smartphone (iPhone), b) by conventional paper-and-pencil questionnaires, and c) by laptop computer. The iPhone and paper versions presented all seven measures. Unfortunately, the computer software available for this study included only four of our seven core measures; thus, we restricted our laptop measures to only those four. To facilitate test-retest reliability measurement of the Mobile Screener content, the iPhone self-assessment was administered for a second time, 1 hour after their first iPhone completion, within the same session, followed finally by usability and satisfaction questions by paper.
Participants and Location
The study participants were 45 active-duty soldiers at a large US military installation. Volunteers were recruited by flyers posted in public areas, and all testing took place at a T2 technology evaluation facility on the installation. All study activities were approved by the installation’s institutional review board.

Measures

Background Questionnaire
Demographic data collected by the paper-and-pencil questionnaire included age, sex, marital status, children, education, race/ethnicity, and military rank. The background questionnaire also required the respondents to rate their preexisting experience and skill with the computer and smartphone technologies.

Screening Measures
The core set of the seven behavioral screening measures in the Mobile Screener was as follows:

1. PTSD Checklist (PCL-C; Weathers and Ford, 1996; Wilkins et al., 2011): The PCL-C is a well-validated self-report measure that evaluates 17 PTSD symptoms using a 5-point Likert-type scale (1 = not at all, 2 = a little bit, 3 = moderately, 4 = quite a bit, and 5 = extremely) with item scores summed (range, 17–85; Weathers, 2008), with higher scores indicating more severe symptoms. Estimates of internal consistency (Cronbach’s alpha) range between 0.94 (Blanchard et al., 1996) and 0.97 (Weathers et al., 1993). Test-retest reliability has been reported as 0.96 at 2 to 3 days and 0.88 at 1 week (Blanchard et al., 1996; Ruggerio et al., 2003).

2. Patient Health Questionnaire–9 (PHQ-9; Kroenke and Spitzer, 2002): The PHQ-9 is a widely used, brief, nine-item, patient self-report depression assessment tool. Items are rated on 4-point scales (0 = not at all, 1 = several days, 2 = more than half the days, 3 = nearly every day) and summed for a possible score ranging from 0 to 27. Higher scores on the PHQ indicate increasing depression severity, with scores higher than 10 reflecting moderate or higher levels of depression (Kroenke et al., 2001). Kroenke et al. (2001) have reported a test-retest reliability of 0.84 and an internal consistency of 0.86 to 0.89.

3. Revised Suicidal Ideation Scale (R-SIS; Rudd, 1989; Rudd and Rajab, 1995): The R-SIS represents a “continuum of suicidal ideation ranging from covert suicidal thoughts to more overt or intense ideation and, ultimately, actual suicide attempts” (Rudd, 1989). The 10 R-SIS items are scored on a 5-point scale (1 = never, 2 = infrequently, 3 = sometimes, 4 = frequently, and 5 = always) and are summed to yield a total ideation score ranging from 10 to 50. The internal consistency cited for the R-SIS has ranged between 0.86 (Rudd, 1989) and 0.91 (Luxtton et al., 2011).

4. Deployment Risk and Resilience Inventory–Unit Support (DRRI-US; King et al., 2006; Vogt et al., 2008): The 12-item DRRI-US assesses assistance and encouragement to service members from unit leaders, peers, and other unit members. Items are scored on a 5-point scale (1 = strongly disagree, 2 = somewhat disagree, 3 = neither agree nor disagree, 4 = somewhat agree, and 5 = strongly agree) and summed for a total support score ranging from 12 to 60. Higher scores are indicative of greater perceived support and cohesion with regard to the military in general, leaders, and fellow unit members. Fikretoglu et al. (2006) have cited an internal consistency for the DRRI-US of 0.91, with a test-retest reliability of 0.78.

5. Dimensions of Anger 5 (DAR5; Hawthorne et al., 2006): The DAR5 has proved to be a reliable, sensitive measure of anger in military populations. The DAR5 is a brief five-item scale with a 5-point response scale (0 = not at all, 1 = a little, 2 = moderately, 3 = a lot, and 4 = very much). Scores are summed to range from 0 to 20, with a higher score indicating greater anger. An internal consistency for the DAR5 of 0.88 has been reported in the literature (Hawthorne et al., 2006).

6. Sleep Evaluation Scale (American Academy of Sleep Medicine Sleep Evaluation Scale; American Academy of Sleep Medicine, 2012): The Sleep Evaluation Scale is a 10-item (true = 1/false = 0) measure developed by the American Academy of Sleep Medicine to evaluate the presence of common sleep difficulties and daytime tiredness. The total score is the sum of all items, with a possible range of 0 to 10. We selected the Sleep Evaluation Scale in part because of our positive experience deploying it as a component of the Afterdeployment.org resource for the military community (Bush et al., 2011).

7. TBI Self-Report of Symptoms: For our brief screening of TBI, we compiled a short series of face-valid items tapping commonly accepted key clinical symptoms. None of the relatively few standardized TBI measures in circulation and other, nonstandardized measures variously used in military clinical practice combined the brevity and the specificity sufficient for delivery by smartphone app. Consequently, we identified common symptom areas from those existing measures and adapted those items for optimum use on a mobile device, in consultation with TBI specialists from the United States Air Force (USAF) TBI Rehabilitation Center and the DoD’s Defense Centers of Excellence for Psychological Health and Traumatic Brain Injury. Our TBI brief screening measure comprises seven (yes = 1/no = 0) items. A user’s endorsement of one or more of these symptoms would alert a provider to pursue further assessment.

Smartphone—User Feedback
Our user feedback questionnaire contained five 10-point items asking the respondents to rate the smartphone content for ease/difficulty of use and likelihood of use. The final two items asked the users about their preferences between the three modes of delivery.

Procedures
Interested service members responded to the study flyers by phoning or e-mailing the study coordinator, who screened individual 90-minute appointments at the T2 facility. The study coordinator greeted each soldier in the assigned test room, explained the study, obtained written informed consent, and assigned a unique study number for use on all measures to ensure anonymity of data. The soldier then completed the background questionnaire followed by the core measures on paper, the laptop computer, and the iPhone, in randomized order. One hour later, the soldier concluded his/her study participation by completing the iPhone measures for a second time, followed by the smartphone usability and satisfaction questions.

Analyses
Our analysis comprised a) calculations of internal consistency by Cronbach’s alpha of all core measures on each medium, b) test-retest reliability using the intraclass correlation coefficient (ICC) of all core iPhone measures between the two iPhone administrations within the test sessions, and c) a test of noninvariance of the scale variances and covariances between the four measures common to all three modes. For the test-retest reliability assessment and between-modality reliability, we chose the ICC because this provides an assessment of absolute agreement in the scores as opposed to an assessment of relative agreement that would be obtained by using Pearson’s correlation coefficient (Streiner and Norman, 2008). We estimated ICCs using Stata, version 12 (STATA: Data Analysis and Statistical Software; STATA, 2012), and we compared the variance-covariance matrix using Mplus (Muthén and Muthén, 2012).
RESULTS

Demographics

Our sample comprised 35 male and 11 female active-duty army (43) or active-duty Army National Guard (2). Sixty percent were married (single/never married, 24%; divorced/legally separated, 16%), and more than half of the sample described themselves as white (white, 52%; African-American, 16%). The soldiers were predominantly junior enlisted (E1–E4, 61%; E5–E9, 32%; two officers and one warrant officer), with only a minority reporting college degrees (General Educational Development/high school diploma or some college, 69%; Associate of Arts degree or higher, 21%).

Preexisting Experience and Skills With Technology

On an ascending scale rating personal-computer skills (novice, advanced beginner, competent, proficient, and expert), 90% judged themselves to be competent or proficient. Most of the participants owned personal smartphones (iPhone or Android, 73%; Blackberry, Windows Mobile, or iPhone plus Blackberry, 16%), and nearly all (95%) rated their skills with smartphones as moderate or advanced (none, beginner, moderate, and advanced). Finally, 41% reported that they used their personal cell phones more than 20 hours a week (6–10 hours per week, 34%; 11–20 hours per week, 18%), most frequently for texting and Internet.

Core Measures

Table 1 depicts the mean total scores and internal consistencies for each of the core measures for each mode of delivery. The Cronbach’s α values throughout for each modality were extremely high, ranging from 0.70 to 0.97, with iPhone figures equivalent to those for the other two modalities. Table 1 also shows the test-retest ICCs for each measure completed on the iPhone. Again, values were high across all measures (ICC, 0.80–0.93). The ICCs between paper, computer, and iPhone, overall, for each core measure were also high throughout (ICC, 0.88–0.92), and estimates for each pair of modalities did not differ appreciably, with the exception of the paper versus computer comparison for the R-SIS, a difference of magnitude but not interpretation. Constraint of the variance-covariance matrix of the PCL-C, the PHQ-9, the RSI-S, and the DRRI-US to equality across the three delivery modes did not result in a statistically significant worsening of model fit ($χ^2(1) = 18.79, df = 20, p = 0.534$), and the comparative fit indices supported the inference of the constrained model as having superior fit to the freely estimated model (Akaike’s information criterion, 3265.21 [free] and 3252.06 [constrained]; Bayes’ information criterion = 3387.23 [free] and 3315.98 [constrained]).

Smartphone—User Feedback

The participants rated ease of use of the iPhone measures on 10-point visual analogue scales, where 1 indicated “very difficult”; and 10, “very easy.” The respondents judged opening and completing the questionnaire items (mean, 9.51; SD, 0.94); submitting answers (mean, 9.64; SD, 0.83); and navigating through the iPhone pages, sections, and questions (mean, 9.44; SD, 1.23) as extremely easy. Similarly, the participants judged themselves as very likely to use the iPhone measures again should the need arise (mean, 8.62; SD, 2.02, where 1 = “not at all likely” and 10 = “very likely.” When asked “if you were going to complete the measures again in the near future, which version (paper, computer, iPhone) would you use?” 73% chose the iPhone, 13% selected the computer, and the remainder had no preference. Finally, 76% of the participants said that they would recommend the iPhone version of the measures to a fellow service member, 11% would recommend the computer-based measures, and 7% had no preference. No participant said that he/she would recommend the paper questionnaire.

© 2013 Lippincott Williams & Wilkins
www.jonmd.com | 993
DISCUSSION

We believe that the results from this study contribute to legitimizing the use of smartphones and similar small-platform mobile devices for collecting self-administered behavioral screening measures. We demonstrated that the reliability of a range of behavioral health questionnaires completed by smartphone was very high, was at least the equivalent of the same measures completed by more conventional means (paper and computer), and compared favorably with values cited in the literature. The psychometric properties of the questionnaires did not seem to be compromised by the small footprint formatting and touch-screen responding of smartphone delivery compared with paper and pencil or keyboard and mouse.

Moreover, the use of the iPhone for questionnaire completion was overwhelmingly preferred by our participants over paper or computer. The volunteers found the interface and the contents easy to navigate and the portable, hip-pocket device convenient.

Limitations

We generated our psychometric data from a relatively modest sample and narrow demographic. In addition, the interval between the test and the retest of our iPhone measures was only 1 hour. Our sample size and test-retest interval were dictated by limited availability and duty restrictions of our military sample. The research literature on this topic would benefit from replication with a larger sample and extended time between iPhone test and retest to counter any practice effects. The literature would also benefit from tests of a wider variety of mobile devices. It is possible that those who volunteered for this study were more technologically knowledgeable and experienced than the army or civilian populations, potentially limiting the generalizability of the results. However, our own research with a considerably larger sample suggests that the personal technology experience, skills, and habits of the typical army service member not only mirror closely this sample but also are characteristic of civilian users (Bush et al., 2012). Finally, our sleep evaluation measure has limited background in the literature, and our brief measure of TBI, although ostensibly face valid, is new and untested. Both of these measures would benefit from more research into their psychometric properties before deployment.

CONCLUSIONS

Our findings show that two fundamental psychometric properties of some widely used self-administered behavioral screening measures delivered by smartphone are highly acceptable for routine use and are equivalent to more conventional modes of administration. We believe that this small but essential study supports the legitimacy of smartphones as convenient and easy-to-use vehicles for behavioral health screening.

DISCLOSURES

This activity was conducted as part of the mission of the National Center for Telehealth and Technology and was not supported by external grant funds.

The authors declare no conflict of interest.

REFERENCES


