Psychological Assessment

Psychometric Properties and Validity of a Mobile Patient Health Questionnaire–9 (MPHQ-9) for Ecological Momentary Assessment in Depressed Adults

Dawson Haddox, Daniel M. Mackin, Tess Z. Griffin, Michael V. Heinz, Matthew D. Nemesure, Amanda C. Collins, George D. Price, Damien Lekkas, Arvind Pillai, Subigya Nepal, Andrew T. Campbell, and Nicholas C. Jacobson

Online First Publication, November 6, 2025. https://dx.doi.org/10.1037/pas0001431

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Assessment

ISSN: 1040-3590

https://doi.org/10.1037/pas0001431

Psychometric Properties and Validity of a Mobile Patient Health Questionnaire–9 (MPHQ-9) for Ecological Momentary Assessment in Depressed Adults

Dawson Haddox[1], Daniel M. Mackin[2], Tess Z. Griffin[3], Michael V. Heinz[2, 3], Matthew D. Nemesure[4], Amanda C. Collins[5, 6], George D. Price[3, 7], Damien Lekkas[8], Arvind Pillai[9], Subigya Nepal[10], Andrew T. Campbell[3, 9], and Nicholas C. Jacobson[3, 9, 11]

[1] Department of Psychology, University of Arizona
[2] Department of Psychiatry, Geisel School of Medicine, Dartmouth College
[3] Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College
[4] Digital Data Design Institute, Harvard Business School, Harvard University
[5] Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, United States
[6] Department of Psychiatry, Harvard Medical School
[7] Quantitative Biomedical Sciences Program, Dartmouth College
[8] Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University
[9] Department of Computer Science, Dartmouth College
[10] Department of Computer Science, University of Virginia
[11] Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College

Ecological momentary assessment is well-suited for capturing rapid symptom dynamics, and it is increasingly used to measure depression symptoms. However, few depression measures are validated for ecological momentary assessment use in the manner expected for traditional questionnaires. Therefore, this study examined the internal consistency, longitudinal stability, and convergent validity of the Mobile Patient Health Questionnaire–9 (MPHQ-9), a version of the Patient Health Questionnaire–9 (PHQ-9) modified for ecological momentary assessment. Depressed participants ($N = 280$; female = 83.93%; White = 79.29%) completed the MPHQ-9 three times daily for 90 days. Data from the first and last 2 weeks were analyzed to align with a prestudy PHQ-9 and poststudy PHQ-9 and Inventory of Depression and Anxiety Symptoms–II. The MPHQ-9 demonstrated fair to substantial adjusted item–total correlations ($r = .42–.83$), often exceeding the PHQ-9 ($r = .39–.72$), with Cronbach's α coefficients of .91 and .81, respectively. Reliability analyses of the MPHQ-9 using generalizability theory and multilevel modeling to account for repeated measures yielded substantial between-person reliability (~1.0) but mixed within-person reliability estimates of .81 (generalizability theory) and .44 (multilevel modeling). The MPHQ-9 showed moderate stability ($r = .69$, intraclass correlation coefficient = .58), compared to the slight stability of the PHQ-9 ($r = .39$, intraclass correlation coefficient = .37). There was moderate agreement between the MPHQ-9 and both the PHQ-9 ($r = .71$) and the Inventory of Depression and Anxiety Symptoms–II General Depression subscale ($r = .65$). Supplementary analyses identified short forms with similar convergent validity but reduced symptom-level information. This study provides initial validation of the MPHQ-9 and compares its psychometric properties to the traditional PHQ-9.

Public Significance Statement

This study provides initial validation of the Mobile Patient Health Questionnaire–9, a version of the widely used Patient Health Questionnaire–9 modified for administration multiple times daily via a smartphone. In a clinical sample from across the United States, the Mobile Patient Health Questionnaire–9 appears to be a valid and reliable measure of depression symptom severity, but further research is required to determine how to interpret its within-person reliability.

Keywords: ecological momentary assessment, depression, Patient Health Questionnaire–9, psychometrics, mobile health

Supplemental materials: https://doi.org/10.1037/pas0001431.supp

Ryan J. Marek served as action editor.

Dawson Haddox https://orcid.org/0009-0004-8595-6454
Daniel M. Mackin https://orcid.org/0000-0002-2188-2968
Tess Z. Griffin https://orcid.org/0000-0001-5462-575X
Michael V. Heinz https://orcid.org/0000-0003-0866-0508
Matthew D. Nemesure https://orcid.org/0000-0002-2369-600X
Amanda C. Collins https://orcid.org/0000-0002-8258-2272
George D. Price https://orcid.org/0000-0002-9164-4973
Damien Lekkas https://orcid.org/0000-0002-6995-9223

continued

Major depressive disorder (MDD) is a prevalent mental health condition characterized by persistent sadness, loss of interest in daily activities, and other psychological and somatic symptoms (American Psychiatric Association, 2013; World Health Organization, 2017). Research indicates that the use of measures to assess depression symptoms can enhance treatment outcomes (Rush, 2015). In clinical and research settings, MDD assessment tools are used to identify the presence and severity of depressive symptoms, guide treatment decisions, and monitor treatment response and symptoms over time, with self-report questionnaires being the most commonly used method.

Among self-report tools, the Patient Health Questionnaire–9 (PHQ-9) stands out for its widespread use and comprehensive validation. This brief questionnaire includes nine items corresponding to each diagnostic criterion for MDD (Kroenke et al., 2001). Respondents retrospectively rate each symptom based on its frequency over the previous 2 weeks. While the PHQ-9 summarizes depression symptoms from the prior 2 weeks, it cannot capture the short-term symptom dynamics that depressed people experience (Crowe et al., 2019). Additionally, retrospective recall may limit reporting accuracy and introduce measurement bias (Ben-Zeev et al., 2009; Horwitz et al., 2023).

Ecological momentary assessment (EMA) has emerged as a promising alternative to address these limitations. EMA involves repeatedly sampling symptoms throughout the day, providing real-time or near-real-time measurements within respondents' naturalistic environments (Aan het Rot et al., 2012; Ebner-Priemer & Trull, 2009; Shiffman et al., 2008). This approach not only provides a more comprehensive and accurate picture of temporal dynamics and symptom fluctuations across short intervals, but it also reduces measurement error compared to traditional retrospective reporting of symptoms. By assessing current or very recent states, EMA aims to reduce recall biases, such as severity or recency biases (Horwitz et al., 2023, 2024), to more accurately assess symptoms (Aan het Rot et al., 2012; Ebner-Priemer & Trull, 2009; Shiffman et al., 2008). Additionally, collecting data in a person's real-world environment enhances ecological validity by measuring symptoms in situations most relevant to them, making findings more generalizable to everyday life. Broadly, EMA can be used to characterize individual differences, describe natural histories, assess contextual associations, and document temporal sequences (Shiffman et al., 2008).

Despite the potential of EMA for studying short-term symptom dynamics and increasing reliability, there remains a lack of validated and agreed-upon EMA measures of MDD. Unlike traditional validated assessments, EMA items and scales are often developed based solely on face validity (Degroote et al., 2020). This approach cannot ensure that items accurately measure their intended constructs. Alternatively, questions are adapted from established retrospective measures by converting the reference period (e.g., asking participants to report on symptoms from the past 4 hr rather than the past 2 weeks; Stone et al., 2023). However, semantic meanings and psychometric properties may change when items are converted from a longer to shorter reference period, necessitating revalidation of the modified items (Stone et al., 2023). For instance, participants asked about anger over the past year remembered only intense episodes, but when asked about the past week, they recalled milder, everyday irritations (Winkielman et al., 1998). Additionally, many EMA scales refer only to a subset of MDD symptoms and may consequently fail to capture the multidimensional and heterogeneous nature of MDD (Dubad et al., 2018).

Few studies have validated specific EMA measures of depression, but some offer insights. Burchert et al. (2021) compared a 14-day EMA using the Moodpath app with the PHQ-9. Participants reported the presence of three out of 17 items corresponding to the *International Classification of Disease*, tenth revision depression symptoms three times daily for 2 weeks. Symptoms marked as present were then rated on a 4-point Likert scale for severity, with clinically significant symptoms (ratings of 3 or 4) counting toward a total score ranging from 0 to 10. The study found moderate agreement between EMA and PHQ-9 scores ($r = .76$, $p < .001$). Targum et al. (2021) examined EMA effectiveness in tracking depression symptoms during an antidepressant trial. For 6 weeks, participants reported on the six items from the $HamD_6$ subscale of the Hamilton Depression Scale, both in the morning (current state) and evening (entire day's state). EMA scores showed fair to moderate consistency with clinician $HamD_6$ ratings at Weeks 2, 4, and 6 ($r = .54$–$.73$, $p < .001$), with gradual increases over time, but no significant correlation at baseline. Jimenez et al. (2022) developed a short-form version of the Dysphoria subscale from the Inventory of Depression and Anxiety Symptoms for EMA use. College students, oversampled for elevated neuroticism, completed EMAs five times daily for 7 days. The Dysphoria EMA Scale showed moderate convergent validity with the baseline

Arvind Pillai https://orcid.org/0000-0002-2489-1130
Subigya Nepal https://orcid.org/0000-0002-4314-9505
Andrew T. Campbell https://orcid.org/0000-0001-7394-7682
Nicholas C. Jacobson https://orcid.org/0000-0002-8832-4741

The authors have no known conflicts of interest to disclose. Funding for the parent study was awarded to Nicholas C. Jacobson by the National Institute of Mental Health and the National Institute of General Medical Sciences, National Institutes of Health (Grant R01MH123482).

Dawson Haddox played a lead role in formal analysis, methodology, and writing–original draft and an equal role in conceptualization, data curation, visualization, and writing–review and editing. Daniel M. Mackin played a supporting role in supervision and an equal role in conceptualization, investigation, methodology, and writing–review and editing. Tess Z Griffin played a lead role in project administration and an equal role in investigation, methodology, and writing–review and editing. Michael V. Heinz played an equal role in investigation, methodology, and writing–review and editing. Matthew D. Nemesure played an equal role in investigation, methodology, and writing–review and editing. Amanda C. Collins played an equal role in investigation, methodology, and writing–review and editing. George D. Price played an equal role in investigation and writing–review and editing. Damien Lekkas played an equal role in data curation, investigation, and writing–review and editing. Arvind Pillai played an equal role in data curation, investigation, software, and writing–review and editing. Subigya Nepal played an equal role in data curation, investigation, methodology, software, and writing–review and editing. Andrew T. Campbell played a lead role in software and supervision and an equal role in investigation, methodology, and writing–review and editing. Nicholas C. Jacobson played a lead role in funding acquisition, investigation, methodology, resources, supervision, and writing–review and editing and an equal role in conceptualization.

Correspondence concerning this article should be addressed to Dawson Haddox, Department of Psychology, University of Arizona, 1503 East University Boulevard (Building 68), Tucson, AZ 85721, United States. Email: dawsonhaddox@arizona.edu

Inventory of Depression and Anxiety Symptoms Dysphoria subscale ($r = .67$, $p < .001$) and demonstrated moderate to substantial internal consistency (omega coefficient = .93 between-person, .68 within-person). An earlier study by Torous et al. (2015) administered an EMA measure based on the PHQ-9 to a small sample of depressed participants three times daily for 29 or 30 days. Each EMA survey included a randomized subset of PHQ-9 items rated on a Likert scale. PHQ-9 estimates from the EMA surveys correlated substantially with traditional PHQ-9 scores ($r = .84$). These studies demonstrate that their EMA depression measures have moderate to substantial convergent validity with traditional assessment tools. Nevertheless, further studies are required to evaluate other EMA designs and psychometric properties beyond convergent validity for depression measures.

EMA studies typically choose between Likert and sliding scales (Haslbeck et al., 2023). A recent preprint compared the two in an EMA design and found that while most psychometric properties showed minimal differences, the sliding scale resulted in larger within-person means and stronger correlations with external criteria related to psychopathology (Haslbeck et al., 2024). Specifically, the sliding scale produced means that were, on average, 30% larger. Additionally, correlations showed model-estimated increases of 77% for the Brief Symptom Inventory and 147%, 44%, and 140% for the Stress, Anxiety, and Depression subscales of the Depression Anxiety Stress Scales. This suggests the sliding scale may be preferable for assessing affective states related to general psychopathology. While it is unclear whether these results generalize to other populations or measures administered via EMA, they may provide the best tentative recommendation until research explores the effects of different response scales for different EMA measures more thoroughly.

The optimal frequency for EMA of depression symptoms remains uncertain and likely depends on the specific symptom and measurement purpose. Symptoms fluctuate over different time frames, with many exhibiting variation within a single day (Crowe et al., 2019; Wichers et al., 2021; Wirz-Justice, 2008). Recent research using signal processing indicates that while symptom fluctuations predominantly occur at slower paces like monthly intervals, measurements at higher frequencies add valuable information by capturing faster, distinct dynamic patterns (Jamalabadi et al., 2024). This finding was consistent across data sets, although the analysis showed a significant peak in signal power at the daily level compared to intraday measurements. Even so, assessing intraday dynamics may often be desirable. For example, when investigating how acute stressors or environmental shifts immediately influence symptoms, intraday assessments can capture transient symptom dynamics that would otherwise be missed. However, practical considerations such as participant burden and compliance rates should be weighed. More frequent assessments might capture rapid changes more precisely, but they might risk overwhelming participants and reducing compliance (S. Wang et al., 2025). On the other hand, some research has failed to find a significant relationship between prompting frequency and compliance (Businelle et al., 2024; Jones et al., 2019; Wrzus & Neubauer, 2023).

Although research has started to validate the psychometric properties of EMA measures of depression, significant gaps remain. Only one study has examined an EMA measure based on the PHQ-9, and that study involved a small sample size of 13 participants (Torous et al., 2015). Additionally, there is a lack of studies that evaluate both convergence with validated measures and other psychometric properties, including internal consistency and long-term stability. Therefore, the present study evaluated the psychometric properties of a Mobile Patient Health Questionnaire–9 (MPHQ-9), an EMA measure of depression adapted from the PHQ-9, in a nationwide clinical sample. The MPHQ-9 preserves the thematic content of the PHQ-9 but alters the time frame to the past 4 hr and adopts a sliding scale response format. The 4-hr reference frame was chosen, based on theoretical rationale due to a lack of empirical consensus, to balance capturing nuanced, intraday fluctuations in symptoms with maintaining manageable participant burden. Specifically, analyses aimed to (a) establish the internal consistency of the MPHQ-9 and compare it with the traditional PHQ-9, (b) evaluate and compare the longitudinal stability of the MPHQ-9 and PHQ-9, and (c) investigate the convergent validity of the MPHQ-9 with the PHQ-9 and the Inventory of Depression and Anxiety Symptoms–Expanded Version (IDAS-II) General Depression subscale.

Method

Transparency and Openness

This study uses data from the *Tracking Depression Study*, funded by the National Institute of Mental Health and the National Institute of General Medical Sciences under Grant R01MH123482. The present analyses were not preregistered. Data will be archived with the National Institute of Mental Health following the embargo period. Analysis code, screenshots of the MPHQ-9, and instructions for training participants to complete the MPHQ-9 are available at the Open Science Framework (Haddox, 2025). The PHQ-9 is publicly accessible, but the IDAS-II is proprietary and not able to be shared by the current authors. See R. Wang et al. (2014) for more information on the MLife/StudentLife app used in the present study.

Participants

The Tracking Depression Study ($N = 300$) is a longitudinal study that used ambulatory assessment methods to capture the behaviors and symptoms of depressed participants over 3 months. Participants were recruited via Meta and Google Ads. Eligibility required residence in the United States, age 18 or older, a current MDD diagnosis, and primary use of an Android smartphone. Exclusion criteria included active suicide risk as determined by the Columbia Suicide Severity Rating Scale and any history of mania or psychosis. Twenty additional participants were excluded from the present analyses for failure to complete postassessments or EMAs spanning dates at least 4 weeks apart, resulting in a present sample of 280 participants. The Dartmouth College Committee for the Protection of Human Subjects approved this study (STUDY00032081).

Most participants in the present sample identified as female (83.93%, $n = 235$), followed by male (11.07%, $n = 31$), nonbinary (3.93%, $n = 11$), and other (1.07%, $n = 3$). Some participants also identified as transgender (3.21%, $n = 9$). In terms of race, the sample identified as primarily White (79.29%, $n = 222$), followed by Black or African American (7.14%, $n = 20$), multiple races (6.43%, $n = 18$), Asian (3.93%, $n = 11$), other races (2.50%, $n = 7$), and American Indian or Alaska Native (0.71%, $n = 2$). Regarding ethnicity, 11.43% ($n = 32$) identified as Hispanic or Latino. At baseline, participants averaged 40.15 years of age ($SD = 11.51$), and 42.50% ($n = 119$) of

participants reported current psychotropic medication use. Additional demographic characteristics are provided in Supplemental Table S1.

Measures

Traditional PHQ-9

The PHQ-9 is a widely used self-report measure of depression symptom severity (Kroenke et al., 2001). The PHQ-9 includes nine items corresponding to each of the diagnostic symptoms of MDD. These include anhedonia, depressed mood, changes in appetite, sleep disturbances, fatigue, feelings of worthlessness or guilt, difficulty concentrating, psychomotor disturbance, and recurrent thoughts of death or self-harm. Participants rate symptom frequency over the past 2 weeks using a 4-point scale from 0 (*not at all*) to 3 (*nearly every day*). The PHQ-9 has demonstrated good internal consistency (Cronbach's $\alpha = .86–.89$), convergent validity ($r = .80$ with the Center for the Epidemiological Studies of Depression–10), and test–rest reliability (0.84 for assessments within 48 hr; Beard et al., 2016; Kroenke et al., 2001).

MPHQ-9

The MPHQ-9 adapts the PHQ-9 for administration via EMA (Torous et al., 2015). The MPHQ-9 maintains PHQ-9 content while modifying the timeframe and response format to capture short-term symptom fluctuations. Participants rate each of the nine depression symptoms from the PHQ-9 using a sliding scale that ranges from 0 (*not at all*) to 100 (*constantly*) over the preceding 4 hr (e.g., "In the past 4 hours, I have had little interest or pleasure in doing things"). In the present study, moving the slider displayed an associated numerical rating from 0 to 100. The default slider position was 50. Research coordinators instructed participants during onboarding to imagine the scale as representing a spectrum from the *least degree* (0) to the *greatest degree* (100) they had ever experienced each symptom. Coordinators further asked participants to maintain a consistent method for responding throughout the study. Figure 1 shows the user interface for the MPHQ-9 in the *Tracking Depression Study*.

IDAS-II

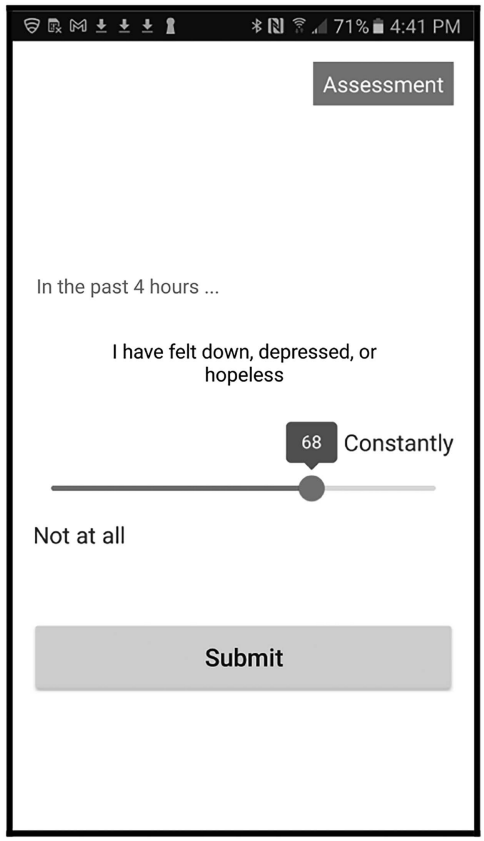
The IDAS-II is a self-report questionnaire designed to assess specific dimensions of depression and anxiety symptoms (Watson et al., 2012). The IDAS-II consists of 99 items mapping onto 18 factor-analytically derived subscales. Respondents rate how much they have experienced each symptom over the past 2 weeks on a 5-point Likert scale ranging from *not at all* to *extremely*, and the items from each subscale are summed. The present study utilized the additional General Depression subscale, which aggregates items from several of the distinct subscales to provide a comprehensive measure of depressive symptomatology. Previous studies have shown that the IDAS-II demonstrates good psychometric properties (Watson et al., 2012).

Procedure

Before enrolling in the study, participants completed the PHQ-9 as part of a screening battery. Subjects who met the enrollment criteria were onboarded and trained for the study procedures. As part of onboarding, participants met with a research coordinator, downloaded the MLife application on their Android devices, received instructions for how to think about each MPHQ-9 EMA throughout the study, reported their wake and sleep times for weekdays and weekends, and completed an example MPHQ-9. Across the 90-day study, MLife prompted participants to complete the MPHQ-9 three times daily, starting with a morning prompt (delivered 4 hr after self-reported wake-up time) and followed by afternoon and evening prompts at 4-hr intervals thereafter. Participants could also complete additional MPHQ-9 questionnaires in the MLife application when desired. Because the MPHQ-9 sleep item (Item 3) was only administered as part of the morning and unprompted additional questionnaires, forward fill imputation was applied to propagate the most recent available rating to the afternoon and evening questionnaires. To enhance compliance, participants received roughly biweekly feedback regarding how many EMAs they had completed via email. Additional compliance emails were sent proactively to participants if they exhibited a decline in compliance or inconsistent adherence. At the end of the study, participants completed a second PHQ-9 and the

Figure 1

Tracking Depression Study User Interface for the Mobile Patient Health Questionnaire–9



Note. The figure shows the user interface for the Mobile Patient Health Questionnaire–9 in the Tracking Depression Study, displaying the depressed mood item (Item 2). Releasing the slider for an item would advance participants to the next item until the questionnaire was completed. Alternatively, participants could change their settings so they only advanced to the next item after pressing the "Submit" button. After progressing, participants could not change their previous ratings. The numerical rating (0–100) appeared when the participant pressed on the slider to move it.

IDAS-II. In total, participants could earn up to $487 based on their completion of all study components.

Data Analytic Strategy

Analyses were conducted in Python v.3.11.8 using each participant's average item and total scores for the first 2 weeks (T1) and the last 2 weeks (T2) of MPHQ-9 data to most closely correspond to the prestudy (T1) PHQ-9 and poststudy (T2) PHQ-9 and IDAS-II. Most analyses focused on T2 data because the interval between the end of MPHQ-9 data collection and the completion of T2 PHQ-9 and IDAS-II was shorter than the interval between the completion of T1 PHQ-9 and the start of MPHQ-9 data collection. To capture within-person variability, supplemental multilevel modeling (MLM) analyses were conducted in R. These analyses used the original sleep item (Item 3) instead of the imputed version, as MLM methods accommodate missing data. Altogether, the analyses focused on internal consistency, longitudinal stability, and convergent validity.

The internal consistency of the T2 MPHQ-9 and T2 PHQ-9 was evaluated using adjusted item–total Pearson correlations, interitem Pearson correlations, and Cronbach's α. The adjusted item–total coefficients were converted into z scores using Fisher's Z transformation and then statistically compared to assess differences between the two assessments (Fisher, 1921). To account for within-person effects, generalizability theory (GT) and MLM approaches for estimating internal consistency reliability were applied to the entire 90-day data set of MPHQ-9 EMAs using the *multilevel.reliability* function from the *psych* R package (Revelle, 2025). GT decomposes variance into person, item, occasion, and two-way interaction components that can be used to estimate between- (Rkr) and within-person (Rc) reliability (Cranford et al., 2006). The MLM approach fits a three-level model—nesting items within occasions within persons—to estimate variance components for each level that are used to compute between- (Rkrn) and within-person (Rcn) reliability coefficients (Nezlek, 2017).

Next, the longitudinal stability of the MPHQ-9 was analyzed by calculating Pearson correlations between T1 and T2 data for each item and the total score. For comparison, Pearson correlations were also calculated between the T1 and T2 PHQ-9. The stability coefficients were also converted into z scores using Fisher's Z transformation and then statistically compared to assess whether stability coefficients differed between the two assessments. To provide an additional measure of stability, reliable change indices (RCs) were calculated for both measures and compared via paired t tests. Further, MLM-derived intraclass correlation coefficients (ICCs) were computed for MPHQ-9 and PHQ-9 item and total scores. Specifically, participants were modeled as random intercepts, with assessment timing (first vs. last 2 weeks) as a binary fixed effect. To evaluate the effect of assessment timing on responses, a standardized correlation was calculated for total scores by dividing the sum of the fixed-effect estimate for assessment timing and its standard deviation by the total score's standard deviation (Rosenthal & Rosnow, 2008).

Finally, convergent validity between the T2 MPHQ-9 and T2 PHQ-9 was assessed by computing Pearson correlations between individual items and scale total scores. Additionally, convergence with the T2 IDAS-II General Depression subscale was examined for both the T2 MPHQ-9 and T2 PHQ-9 by computing Pearson correlations between corresponding items and scale total scores. These Pearson correlation coefficients were converted into z scores using Fisher's Z transformation and then statistically compared to assess whether the degree of convergence with the IDAS-II differed between the T2 MPHQ-9 and T2 PHQ-9. An attrition analysis was also conducted using demographic variables, T1 PHQ-9 scores, and T1 and T2 MPHQ-9 scores to determine whether participants who were included in the study differed significantly from those who were excluded.

Results

Descriptive Statistics

Participants were prompted to complete 42 MPHQ-9 EMAs at both T1 (first 2 weeks of study) and T2 (last 2 weeks of study), with an average completion of 37.51 EMAs ($SD = 9.09$) at T1 and 33.10 EMAs ($SD = 9.24$) at T2. They completed one or more MPHQ-9 EMAs on an average of 12.37 days ($SD = 2.18$) at T1 and 12.44 days ($SD = 2.47$) at T2. The average time between the completion of the T1 PHQ-9 and the study's start was 60.52 days ($SD = 37.22$) due to the Tracking Depression Study enrollment protocol. Participants completed the T2 PHQ-9 and T2 IDAS-II an average of 2.38 days ($SD = 5.11$) after the study ended. An attrition analysis of all demographic variables, T1 PHQ-9 scores, and T1 and T2 MPHQ-9 scores found significant differences between included and excluded participants for sexual orientation, $\chi^2(N = 280) = 16.81$, $p = .02$, and the T2 MPHQ-9 suicide item, $t(300) = -2.51$, $p = .01$. Specifically, adjusted residuals indicated that bicurious participants had a higher rate of exclusion than expected (adjusted residual = 3.71, $|z| > 1.96$), and excluded participants reported higher scores on the T2 MPHQ-9 suicide item. Table 1 presents the means and standard deviations for MPHQ-9 and PHQ-9 scores.

Internal Consistency

As shown in Table 2, adjusted item–total correlations ranged from fair to substantial, $r(280) = .42–.83$, $ps < .001$, for the T2 MPHQ-9 and slight to moderate, $r(280) = .39–.72$, $ps < .001$, for the T2 PHQ-9 (Shrout, 1998). Fisher's Z-transformation tests revealed that the T2 MPHQ-9 had significantly higher adjusted item–total correlations than the T2 PHQ-9 for all items ($zs = -5.39$ to -2.05, $ps < .001$ for Items 1–7, $p < .05$ for Item 8) except for suicidal ideation ($z = -0.09$, $p = .93$). Interitem correlations for the T2 MPHQ-9, also shown in Table 2, ranged from slight to substantial, $r(280) = .26–.87$, $ps < .001$, with a fair average of $r = .51$. For the T2 PHQ-9, interitem correlations ranged from slight to moderate, $r(280) = .11–.71$, with a slight average of $r = .32$, and they were all significant except for the correlation between changes in appetite and suicidal ideation. Cronbach's α coefficients were .91 for the T2 MPHQ-9 and .81 for the T2 PHQ-9, indicating high internal consistency for both measures.

The between-person reliability coefficients from both GT and MLM rounded to 1.0, but within-person reliability was higher using GT (Rc = .81) than MLM (Rcn = .44). Prior EMA research has similarly found lower within-person reliability for MLM than GT (Castro-Alvarez et al., 2024), though the discrepancy observed here is more pronounced. These findings suggest that the MPHQ-9 demonstrates strong reliability when used to compare individuals' average scores but that within-person reliability depends on whether time and items are modeled as crossed or nested factors. See

Table 1
Means and Standard Deviations for MPHQ-9 and PHQ-9 Scores

Questionnaire item	MPHQ-9		PHQ-9	
	First 2 week (T1)	Last 2 week (T2)	Prestudy (T1)	Poststudy (T2)
1. Anhedonia	50.12 ± 18.26	52.23 ± 23.69	2.44 ± 0.70	2.03 ± 0.90
2. Depressed mood	50.32 ± 19.23	49.48 ± 24.83	2.49 ± 0.66	1.95 ± 0.96
3. Sleep difficulty	51.56 ± 20.28	51.34 ± 24.04	2.30 ± 0.89	1.92 ± 0.99
4. Fatigue	63.59 ± 17.42	64.32 ± 20.42	2.69 ± 0.56	2.45 ± 0.77
5. Changes in appetite	43.72 ± 21.35	43.77 ± 26.26	1.97 ± 0.96	1.55 ± 1.04
6. Worthlessness and guilt	50.41 ± 21.97	49.95 ± 26.63	2.26 ± 0.89	1.84 ± 1.02
7. Concentration	53.59 ± 21.10	52.42 ± 24.19	1.99 ± 0.96	1.67 ± 1.03
8. Psychomotor disturbance	28.68 ± 21.39	28.93 ± 24.61	0.59 ± 0.83	0.64 ± 0.81
9. Suicidal ideation	10.10 ± 15.81	8.77 ± 15.81	0.50 ± 0.72	0.38 + 0.69
Total	402.08 ± 128.92	401.19 ± 160.23	17.21 ± 3.65	14.44 ± 5.19

Note. To ensure each participant's data contributed equally to the overall metrics, the means and standard deviations for the MPHQ-9 were calculated from the T1 and T2 average scores for each participant, rather than for the raw scores. MPHQ-9 = Mobile Patient Health Questionnaire–9; PHQ-9 = Patient Health Questionnaire–9.

Supplemental Tables S2–S3 for the variance components obtained through GT and MLM, which were used to calculate the reliability coefficients.

Stability

Table 3 presents the longitudinal stability of scores from the traditional and mobile versions of the PHQ-9. The stability coefficients were slight for the PHQ-9, rs(280) = .30–.41, ps < .001, and moderate for the MPHQ-9, rs(280) = .59–.82, ps < .001. Fisher's Z-transformation comparisons showed that the MPHQ-9's stability coefficients were significantly higher across all items and the total score, indicating greater reliability over time compared to the PHQ-9 (zs = −9.97 to −2.85, ps < .001).[1] RCs for MPHQ-9 and PHQ-9 scores from T1 to T2, along with ICCs for the MPHQ-9 and PHQ-9 categorized by completion timing (T1 or T2), are also in Table 3. The RCs differed significantly for Items 1 through 7 and the total score, with MPHQ-9 scores showing less change (p < .001 for Items 1 through 6 and the total score, p < .01 for Item 7). The ICCs ranged from .38 to .67 for the MPHQ-9 and from .28 to .41 for the PHQ-9. The fixed effect of assessment timing (T1 or T2) on the MPHQ-9 was not significant (β = −1.86, SE = 1.59), t(19508) = −1.17, p = .24. The standardized coefficient (Rosenthal & Rosnow, 2008) was −.008, suggesting a negligible relationship. For the PHQ-9, the fixed effect was significant (β = −2.78, SE = 0.30), t(279) = −9.24, p < .001, and the standardized coefficient was −.48, indicating a fair relationship. Short-term stability metrics, including root-mean-square of successive differences values and autocorrelations, are provided for T2 MPHQ-9 scores in Supplemental Tables S4–S5.

Convergent Validity

Pearson correlations between the PHQ-9 and MPHQ-9 are displayed in Table 4. Matching-item correlations ranged from fair to moderate in magnitude, rs(280) = .53–.70, ps < .001. The correlation between the T2 PHQ-9 and T2 MPHQ-9 total scores was moderate, r(280) = .71, p < .001, suggesting a reasonable level of agreement between the two measures.[2]

Pearson correlations were also calculated between the T2 MPHQ-9 and T2 PHQ-9 scores and the corresponding item and total scores for the T2 IDAS-II General Depression subscale (see Table 5). The matching-item correlations for the T2 MPHQ-9 and IDAS-II ranged from slight to moderate, r(280) = .37–.69, ps < .001, and the total score correlation was moderate, r(280) = .65, p < .001, indicating a reasonable level of agreement. Matching-item correlations between the PHQ-9 and IDAS-II also ranged from slight to moderate, r(280) = .33–.78, ps < .001. The PHQ-9 IDAS-II total score correlation was substantial, r(280) = .83, p < .001, suggesting high agreement between the measures.[3]

Fisher's Z-transformation tests comparing the magnitude of corresponding MPHQ-9 IDAS-II and PHQ-9 IDAS-II matching-item correlations and total correlations are reported in Table 5. Results indicated that the correlations between the PHQ-9 and the IDAS-II were significantly larger for the total score and all items except fatigue and changes in appetite (zs = 2.04–4.71; ps < .001 for anhedonia, suicidal ideation, and the total score; p < .01 for worthlessness and guilt; and ps < .05 for depressed mood, sleep difficulty, concentration, and psychomotor disturbance).

Discussion

The present study investigated the psychometric properties of the MPHQ-9 to determine its appropriateness for use in EMA studies. Specifically, this study examined the internal consistency, stability, and convergent validity of the MPHQ-9, an EMA measure of depression based on the PHQ-9, as compared to the traditional PHQ-9. Results showed that the MPHQ-9 demonstrated good psychometric properties, and it exhibited higher internal consistency and stability compared to the PHQ-9.

[1] Partial correlations accounting for the varying time gap between T1 and T2 PHQ-9 completion were identical with regard to the pattern of significance and strength.

[2] Partial correlations accounting for occasional time gaps between the end of T2 MPHQ-9 data collection and completion of the T2 PHQ-9 were identical with regard to the pattern of significance and strength.

[3] Partial correlations accounting for occasional time gaps between the end of T2 MPHQ-9 data collection and completion of the T2 IDAS-II General Depression subscale were identical with regard to the pattern of significance and strength.

Table 2
Adjusted Item–Total and Interitem Correlations for the MPHQ-9 and PHQ-9

Questionnaire item	MPHQ-9: *r*	PHQ-9: *r*	Interitem correlation								
			1	2	3	4	5	6	7	8	9
1. Anhedonia	.81***[a]	.68***	—	.71	.32	.52	.29	.45	.42	.31	.32
2. Depressed mood	.83***[a]	.72***	.85	—	.33	.49	.30	.61	.38	.31	.38
3. Sleep difficulty	.63***[a]	.39***	.52	.51	—	.36	.28	.18	.19	.17	.19
4. Fatigue	.76***[a]	.54***	.78	.69	.65	—	.26	.28	.36	.23	.25
5. Changes in appetite	.60***[a]	.39***	.47	.48	.49	.47	—	.25	.23	.27	.11
6. Worthlessness and guilt	.77***[a]	.51***	.70	.87	.44	.59	.52	—	.31	.16	.37
7. Concentration	.76***[a]	.49***	.69	.68	.52	.68	.45	.66	—	.43	.25
8. Psychomotor disturbance	.55***[a]	.42***	.42	.44	.43	.39	.48	.41	.53	—	.29
9. Suicidal ideation	.42***	.42***	.36	.40	.29	.26	.28	.39	.34	.33	—

Note. Values were computed using the last two weeks of MPHQ-9 data (T2) and the poststudy PHQ-9 (T2). The interitem correlations are shown in the bottom-left of the correlation matrix for the MPHQ-9 and in the top-right for the PHQ-9. MPHQ-9 = Mobile Patient Health Questionnaire–9; PHQ-9 = Patient Health Questionnaire–9.
[a] The adjusted item–total correlation for this item is significantly higher than on the other PHQ-9 version ($p < .05$). All MPHQ-9 interitem correlations were significant ($ps < .001$). All PHQ-9 interitem correlations were significant, except between Items 5 and 9 ($p = .064$). Most were significant at $p < .001$, except for between Items 6 and 8 and for Item 3 with Items 6, 7, 8, and 9 ($ps < .01$).
*** $p < .001$.

Main Findings

The item–total correlations for the T2 MPHQ-9 ranged from fair to substantial, while those for the T2 PHQ-9 ranged from slight to moderate. The T2 MPHQ-9 had significantly higher item–total correlations for all items except the one assessing suicidal ideation. Interitem correlations for the T2 MPHQ-9 ranged from slight to substantial with a fair average, while for the T2 PHQ-9, they ranged from slight to moderate with a slight average. Cronbach's α coefficients were substantial for the T2 MPHQ-9 and the T2 PHQ-9, but they were higher for the T2 MPHQ-9. These results suggest that the MPHQ-9 exhibits higher internal consistency than the PHQ-9. While prior studies have reported slightly higher correlations (mean interitem = .55, mean item–total = .70; Choi et al., 2014) and Cronbach's α coefficients (α = .86–.91) for the PHQ-9 (Beard et al., 2016; Bianchi et al., 2022; Choi et al., 2014; Kroenke et al., 2001), these values were still lower or comparable to those found for the T2 MPHQ-9. The discrepancy in performance of the T2 MPHQ-9 suicide item compared to the other items on the scale may stem from the exclusion of subjects at active suicide risk at T1.

To account for repeated measures, GT and MLM approaches were used to estimate between- and within-person reliability. Both methods found high between-person reliability, which indicates that persons' mean scores were measured with minimal error. In psychological dynamics research, these means are often interpreted as trait scores, representing the stable aspects of psychological constructs (Nezlek, 2017; Castro-Alvarez et al., 2024). High between-person reliability is crucial for supporting research on long-term, enduring changes.

Table 3
Stability of MPHQ-9 and PHQ-9

Questionnaire item	MPHQ-9			PHQ-9		
	Mean RC	*r*	ICC	Mean RC	*r*	ICC
1. Anhedonia	0.14[a]	.64***[a]	.45 (.40–.49)	−0.50	.32***	.31 (.20–.41)
2. Depressed mood	−0.05[a]	.67***[a]	.48 (.44–.52)	−0.69	.30***	.28 (.16–.38)
3. Sleep difficulty	−0.01[a]	.59***[a]	.38 (.34–.42)	−0.38	.41***	.41 (.30–.49)
4. Fatigue	0.05[a]	.64***[a]	.40 (.36–.44)	−0.38	.35***	.33 (.23–.42)
5. Changes in appetite	0.00[a]	.66***[a]	.48 (.44–.52)	−0.40	.39***	.39 (.29–.48)
6. Worthlessness and guilt	−0.03[a]	.68***[a]	.55 (.50–.59)	−0.43	.41***	.40 (.30–.50)
7. Concentration	−0.07[a]	.70***[a]	.51 (.47–.55)	−0.30	.39***	.39 (.30–.49)
8. Psychomotor disturbance	0.02	.70***[a]	.56 (.52–.60)	0.06	.41***	.41 (.31–.50)
9. Suicidal ideation	−0.14	.82***[a]	.67 (.63–.70)	−0.14	.30***	.30 (.19–.40)
Total	−0.01[a]	.69***[a]	.58 (.54–.62)	−0.69	.39***	.37 (.26–.47)

Note. The T1–T2 gap for the MPHQ-9 (calculated from the first 2 weeks [T1] to the last 2 weeks [T2] of MPHQ-9 data collection) averaged 89.30 days ($SD = 3.50$) from the start of T1 data collection to the end of T2 data collection, while the gap for the traditional PHQ-9 (calculated from the prestudy [T1] PHQ-9 to poststudy [T2] PHQ-9) averaged 153.25 days ($SD = 37.48$). ICC values are reported as ICC (95% confidence interval), where the parentheses contain the confidence interval. MPHQ-9 = Mobile Patient Health Questionnaire–9; PHQ-9 = Patient Health Questionnaire–9; RC = reliable change index; ICC = intraclass correlation coefficients.
[a] The T1–T2 correlation or RC for this item is significantly higher than for the corresponding item on the other PHQ-9 version.
*** $p < .001$.

Table 4
Agreement Between MPHQ-9 and PHQ-9 Scores

Questionnaire item	*r*
1. Anhedonia	.62***
2. Depressed mood	.70***
3. Sleep difficulty	.52***
4. Fatigue	.55***
5. Changes in appetite	.70***
6. Worthlessness and guilt	.66***
7. Concentration	.61***
8. Psychomotor disturbance	.54***
9. Suicidal ideation	.68***
Total	.71***

Note. Values were computed using the last two weeks of MPHQ-9 data (T2) and the poststudy PHQ-9 (T2). MPHQ-9 = Mobile Patient Health Questionnaire–9; PHQ-9 = Patient Health Questionnaire–9.
*** $p < .001$.

However, while GT estimated substantial within-person reliability for the MPHQ-9, MLM indicated only fair reliability, with slightly less than half of the observed within-person fluctuation deemed reliable. These within-person reliability coefficients capture the consistency of persons' responses at a given occasion (Nezlek, 2017) and the accuracy in reflecting true intraindividual variability (Cranford et al., 2006; Neubauer & Schmiedek, 2020). Ensuring adequate within-person reliability is crucial to distinguish meaningful fluctuations from random error. Unfortunately, methodological research on how to interpret the reliability of a measure when these coefficients differ is lacking.

Overall, these findings suggest that GT's crossed design—where person, item, and time are overarching factors—explains more variance than MLM's nested design, which treats time as person-specific and items as nested within person–time. Although both methods estimate within-person reliability as the ratio of "true" to total variance, their variance components differ. GT defines this ratio as the variance in person–time interactions over the sum of that variance and measurement error (three-way person–time–item interactions) adjusted for item count. MLM uses time-level variance over the sum of time- and item-level variance adjusted for item count. A key distinction lies in their treatment of item- and time-related variance. GT treats overarching item effects (e.g., item difficulty across persons and time) and two-way interactions as meaningful, considering only three-way interactions as error. MLM assumes items exist solely within person–time, preventing it from separating overarching item effects or two-way interactions and instead attributing all item-related variance to error. Similarly, GT differentiates overarching time effects and their two-way interactions, whereas MLM assumes persons have separate time universes. The choice between GT and MLM partly reflects whether these separations seem reasonable and whether systematic differences are considered true signal or error. Importantly, some item effects—such as item difficulty or person-specific symptom susceptibility—may influence ratings without affecting within-person consistency. In this data set, GT identified item effects and person–item interactions as major variance sources in MPHQ-9 responses (25% and 19%, respectively), while item–time interactions were negligible (approximately 0%). Future research should formally investigate how to interpret variance components and coefficient divergence.

Table 5
Agreement Between the MPHQ-9 and PHQ-9 With the IDAS-II

PHQ-9 item	IDAS-II item	MPHQ-9 *r*	PHQ-9 *r*
1. Anhedonia	2, 27	.57***	.74***[b]
2. Depressed mood	8	.69***	.77***[b]
3. Sleep difficulty	11, 51	.61***	.71***[b]
4. Fatigue	6, 64	.50***	.57***
5. Changes in appetite	1, 26	.37***	.33***
6. Worthlessness and guilt	21, 31	.57***	.70***[b]
7. Concentration	61	.62***	.71***[b]
8. Psychomotor disturbance	5, 57	.47***	.60***[b]
9. Suicidal ideation	13, 52	.59***	.78***[b]
Total	General depression[a]	.65***	.83***[b]

Note. Values were computed using the last two weeks of MPHQ-9 data (T2) and the poststudy PHQ-9 and IDAS-II (T2). MPHQ-9 = Mobile Patient Health Questionnaire–9; PHQ-9 = Patient Health Questionnaire–9; IDAS-II = Inventory of Depression and Anxiety Symptoms–Expanded Version.
[a] IDAS-II Items 9, 30, 40, and 48 were excluded from correlations with specific PHQ-9 items but included in correlations between PHQ-9 version totals and the IDAS-II General Depression subscale. [b] The item's correlation with the IDAS-II is significantly higher than the corresponding item on the other PHQ-9 version's correlation with the IDAS-II.
*** $p < .001$.

Because item–total and interitem correlations based on participants' mean scores only reflect between-person trends, we also computed within-person correlations (see Supplemental Material 5). These analyses revealed marked heterogeneity across individuals. Within-person adjusted item–total correlations were generally fair, with mean values ranging from .18 ($SD = 0.28$; suicidal ideation, sleep difficulty) to .52 ($SD = 0.25$; depressed mood). Within-person interitem correlations were more variable, spanning trivial to moderate magnitudes, with means ranging from .04 ($SD = 0.23$; sleep–suicidal ideation) to .53 ($SD = 0.31$; anhedonia–depressed mood), again accompanied by substantial variability across individuals.

The PHQ-9 showed only slight stability from T1 to T2, whereas the MPHQ-9 exhibited moderate stability. Notably, the MPHQ-9 demonstrated significantly greater stability across all items and the total score, as evidenced by both higher Pearson correlations between T1 and T2 scores and smaller RC absolute values for item and total scores. This indicates that the MPHQ-9 maintained greater consistency over time compared to the PHQ-9. MLM-derived ICCs for the MPHQ-9 were mostly fair, and T1/T2 assessment timing did not meaningfully predict MPHQ-9 scores. In contrast, the PHQ-9 showed mostly slight ICCs, with assessment timing as a fair predictor. Given the episodic nature of MDD, some symptom fluctuation is expected. However, no treatments were administered as a part of the study, so a degree of stability was also anticipated. Future research should explore whether the PHQ-9's lower stability reflects measurement error or greater sensitivity to actual symptom changes.

The matching-item correlations between the T2 MPHQ-9 and T2 PHQ-9 ranged from fair to moderate, with a moderate correlation between the total scores. Similarly, the T2 MPHQ-9 and T2 IDAS-II General Depression subscale matching-item correlations ranged from slight to moderate, and total scores were moderately correlated. This indicates that the MPHQ-9 has reasonable external validity as an EMA measure of depression symptoms. When comparing the magnitude of the IDAS-II General Depression correlations with the MPHQ-9 versus the T2 PHQ-9, correlations between the T2 PHQ-9

were significantly larger for the total score and all items except fatigue and changes in appetite. Taken together, these results are similar to previous studies looking at the convergence between EMA and retrospective measures of depression symptoms (Burchert et al., 2021; Jimenez et al., 2022; Targum et al., 2021). As for why the MPHQ-9 demonstrates weaker convergence with the IDAS-II than the PHQ-9, the PHQ-9 and IDAS-II are both retrospective measures, employ Likert scales, and were completed within minutes of each other. Overall, while the MPHQ-9 showed strong convergent validity with established retrospective depression measures, the PHQ-9 demonstrates even stronger convergent validity.

Several reasons may explain the differences between the psychometric properties of the MPHQ-9 and PHQ-9. First, switching from a Likert to a sliding scale might have introduced a method effect. The 101-point sliding scale used in the MPHQ-9 can capture slight differences in symptoms, which minimally affect the total score, compared to the PHQ-9's 4-point Likert scale, where a 1-point change in an item has a much larger relative effect on the total score.

Second, the MPHQ-9 is an EMA measure, while the PHQ-9 is a retrospective measure. Thus, the MPHQ-9 is administered in participants' naturalistic environments and involves a shorter recall window (4 hr vs. 2 weeks), potentially reducing recall bias. Further, averaging multiple MPHQ-9 scores for each participant could reduce within-person error. It is also possible that items are interpreted differently in a momentary versus retrospective measure, and momentary measures may be more likely to trigger careless responding. Indeed, retrospective measures may not be ideal for assessing construct validity for EMA (Dubad et al., 2018). As such, future research should employ different designs to explore reasons for the differences in convergence between the MPHQ-9 and PHQ-9, such as whether the differences result from the superior construct validity of the PHQ-9 or the MPHQ-9.

Third, the timings of MPHQ-9 and PHQ-9 completion were misaligned. The T2 PHQ-9 was completed on average 2.97 days ($SD = 5.29$) after the end of T2 MPHQ-9 data collection. Additionally, the average time gap between T1 and T2 was shorter for the MPHQ-9 than the PHQ-9, as the PHQ-9 was administered before and after EMA data collection, while the MPHQ-9 was completed during the first and last 2 weeks of EMA data collection (see Procedure). These misalignments may have affected convergent validity and internal consistency. However, secondary analyses controlling for time demonstrated an identical pattern of findings (see Footnotes 1–3).

Last, the MPHQ-9 and PHQ-9 ask about slightly different constructs. The PHQ-9 asks about how many days each symptom was experienced over the past 2 weeks, while the 2-week MPHQ-9 averages aggregate many MPHQ-9s asking how often symptoms were experienced over the prior 4 hr. Additionally, participants were advised to view the MPHQ-9 extremes as representing the least to the greatest degree to which they had ever experienced each symptom, meaning that participants may have also considered symptom intensity rather than just symptom frequency. These are connected but distinct aspects of symptom severity. To summarize, the MPHQ-9 may have demonstrated different psychometric properties because it was administered via EMA rather than as a retrospective measure, used a sliding rather than Likert scale, had a misalignment in the timing of T1 and T2 completion compared to the PHQ-9, and asked about slightly different constructs than the PHQ-9.

EMA requires balancing measurement comprehensiveness with participant burden. High interitem correlations in the T2 MPHQ-9 scores suggest potential redundancy when creating sum scores. We derived three sets of MPHQ short forms using an all-subsets item selection approach with five-fold outer and three-fold inner cross-validation. These sets of short forms were separately optimized to predict the IDAS-II General Depression subscale total, the PHQ-9 total, and the momentary MPHQ-9 total. We identified viable short forms of varying lengths. As an example, four-item versions provided a reasonable balance between performance and length, demonstrating strong convergence with the full nine-item scale. Across participants' T2 averages, correlations with the nine-item total were $r = .96$–$.97$, mean within-person correlations (averaged in Fisher Z space) were $r = .89$–$.91$, external convergence was comparable to the full MPHQ-9, and internal consistency remained substantial ($\alpha = .84$–$.85$ vs. .91 for MPHQ-9). Complete procedures appear in Supplemental Material 4.

Despite the utility of short forms for aggregate scoring, individual items retain unique information that is lost when items are removed. We conducted regression analyses to quantify this unique variance using two approaches. First, we used person-mean-centered EMA data in fixed-effects ordinary least square models, predicting each item from the remaining items to isolate within-person variance patterns. Second, we used T2 average scores in standard ordinary least square models to examine between-person variance patterns. Momentary items retained 53%–94% unexplained variance in within-person analyses after controlling for other items. When using T2 MPHQ-9 averages, unexplained variance ranged from 13% to 79%. Hierarchical ordinary least squares regression models examined clinical meaningfulness using IDAS-II symptom composites as outcomes (defined using the mapping from Table 5). MPHQ-9 total scores were entered first, followed by corresponding individual items. Individual items significantly improved captured variance beyond total scores (incremental $R^2 = .04$–$.25$, all $p < .001$). These findings demonstrate that while shortened scales efficiently provide global depression scores, the full nine-item MPHQ-9 captures important information for research examining specific depressive symptoms. Detailed results appear in Supplemental Material 6.

Implications for Research and Practice

The findings of the present study have implications for both research and practice. First, they provide preliminary evidence that the MPHQ-9 is a valid EMA measure of depression. This is significant given the scarcity of validated EMA measures of depression and the risks associated with relying on face validity or untested adaptations of retrospective measures. Future EMA research should prioritize using validated EMA measures when possible, and the MPHQ-9 is now one of only a few evidence-based options available. Second, the findings suggest that the MPHQ-9 possesses different psychometric properties when administered to depressed populations compared to the widely used retrospective PHQ-9. The aggregated MPHQ-9 exhibited higher reliability compared to the PHQ-9 and also showed strong convergent validity. Theoretical justifications propose that momentary measures like the MPHQ-9 are less prone to recall biases and capture symptoms in higher resolution. This allows researchers and clinicians to measure not only a summary of how often symptoms were experienced over the past 2 weeks but also the symptoms experienced at specific times throughout the day and the variability in symptoms across days. Nonetheless, deciding between using

retrospective and momentary measures to assess depression symptoms should involve a context-based consideration of pros and cons as well as how retrospective and momentary measures might be used in tandem. While the MPHQ-9 captures symptoms in higher resolution, it also places a higher burden on the person being assessed. Further, the within-person reliability of the MPHQ-9 remains questionable due to the inconsistent results provided by the GT and MLM approaches. The between-person reliability of the MPHQ-9 is high across approaches, but shorter measures could likely yield similar between-person reliability of means with fewer items and thereby reduce respondent burden. Moreover, although the MPHQ-9 may better capture experienced symptoms, broader recollections might better predict certain behaviors (Kahneman et al., 1993; Redelmeier et al., 2003; Stone et al., 2023). Additionally, future research might consider how retrospective and momentary measures can provide complementary and additive insights in research and practice.

Strengths and Limitations

This study has several strengths that enhance its significance for EMA of depression. It is among the first to evaluate the psychometric properties of an EMA measure of depression using methods expected for traditional clinical questionnaires. The use of a nationally recruited clinical sample of depressed participants also increases the generalizability of the findings to the broader depression population in the United States. Additionally, the study assessed both the internal and external properties of the MPHQ-9, expanding on previous articles that have typically focused on one aspect. Furthermore, by examining the convergence between the MPHQ-9 and not only one but two established measures—the PHQ-9 and IDAS-II—this study provides a more robust validation. This dual comparison reduces the risk of bias associated with relying on a single comparator and more accurately reflects the measure's alignment with established tools.

The present study also has several limitations. First, the timing of the EMA and retrospective assessments was not completely aligned. Even so, analyses controlling for the difference in the timing of these assessments showed this difference did not impact the pattern of results. Second, while the findings suggest that the aggregated MPHQ-9 may reliably assess depression symptoms, there is still uncertainty about how to interpret its scores, particularly in linking them to specific severity levels. Nonetheless, it is at least clear that the MPHQ-9 is measuring depression symptoms with some level of reliability and validity, and MPHQ-9 scores can be compared and interpreted relative to other MPHQ-9 scores. Third, participants were asked to interpret scale extremes based on their past experience and to use a consistent approach. This may have improved within-person standardization but did not address between-person standardization. Despite this, the MPHQ-9 exhibited high between-person reliability coefficients according to the GT and MLM approaches. Last, the MPHQ-9's sliding scale defaulted to a rating of 50, and prior research suggests that default values can bias responses to 101-point sliding scales (Liu & Conrad, 2019). Removing default values may improve the performance of the measure while also increasing generalizability to nonclinical populations.

Constraints on Generality

Several factors may limit the generalizability of these findings. First, the psychometric properties of the MPHQ-9 in nondepressed populations remain uncertain. Additionally, the exclusion of individuals at risk of suicide or with a history of mania or psychotic symptoms restricts the applicability of the results to important subgroups within the broader mood disorder spectrum. Second, the sample was predominantly White (79.29%), cisgender (96.79%), heterosexual (66.43%) women (83.93%) with at least some college education (93.28%), primarily residing in the United States and using Android smartphones. Thus, the findings may not generalize to populations with different gender identities, racial and ethnic backgrounds, socioeconomic statuses, or cultural experiences. Although efforts were made to recruit a nationally representative sample of treatment-seeking adults, future studies should prioritize more diverse participant groups to evaluate MPHQ-9 invariance across subpopulations. Last, the study involved three daily assessments, and participants did not receive study-related treatment. Therefore, generalizability to other EMA schedules or treatment contexts remains unknown. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

Conclusion

In conclusion, this study provides preliminary evidence for the validity and reliability of the MPHQ-9 as an EMA measure of MDD. The MPHQ-9 demonstrated strong internal consistency, long-term stability, and convergent validity, with higher internal consistency and long-term stability compared to the PHQ-9. This study is one of few to propose a measure of depression validated for use in EMA designs. Using validated assessments like the MPHQ-9 would allow future studies to have greater confidence in their measurements, while standardizing validated EMA depression measures would facilitate comparison of results across different EMA studies. Nonetheless, future research should also continue to refine the methods for assessing the psychometric properties and validity of EMA measures, and new and revised measures based on these findings should continually be developed. Future research concerning the MPHQ-9 should investigate the psychometric properties of the MPHQ-9 in other samples, explore reasons for the lower within-person reliability coefficient under MLM compared to GT, develop guidelines for its use in diagnosing and assessing symptom severity, evaluate how changes to the scale based on best practices (e.g., removing the default slider value, removing the ambiguity in whether instructions refer to frequency or intensity) affect the measure's psychometric properties, determine the reasons for the differences in the MPHQ-9's and PHQ-9's psychometric properties, and consider whether and how retrospective and momentary measures can provide unique and complementary benefits.

References

Aan het Rot, M., Hogenelst, K., & Schoevers, R. A. (2012). Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies. *Clinical Psychology Review*, *32*(6), 510–523. https://doi.org/10.1016/j.cpr.2012.05.007

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

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. https://doi.org/10.1016/j.jad.2015.12.075

Ben-Zeev, D., Young, M. A., & Madsen, J. W. (2009). Retrospective recall of affect in clinically depressed individuals and controls. *Cognition and Emotion*, *23*(5), 1021–1040. https://doi.org/10.1080/02699930802607937

Bianchi, R., Verkuilen, J., Toker, S., Schonfeld, I. S., Gerber, M., Brähler, E., & Kroenke, K. (2022). Is the PHQ-9 a unidimensional measure of depression? A 58,272-participant study. *Psychological Assessment*, *34*(6), 595–603. https://doi.org/10.1037/pas0001124

Burchert, S., Kerber, A., Zimmermann, J., & Knaevelsrud, C. (2021). Screening accuracy of a 14-day smartphone ambulatory assessment of depression symptoms and mood dynamics in a general population sample: Comparison with the PHQ-9 depression screening. *PLOS ONE*, *16*(1), Article e0244955. https://doi.org/10.1371/journal.pone.0244955

Businelle, M. S., Hébert, E. T., Shi, D., Benson, L., Kezbers, K. M., Tonkin, S., Piper, M. E., & Qian, T. (2024). Investigating best practices for ecological momentary assessment: Nationwide factorial experiment. *Journal of Medical Internet Research*, *26*, Article e50275. https://doi.org/10.2196/50275

Castro-Alvarez, S., Zhou, D. J., Bringmann, L., Tutunji, R., Proppert, R. K. K., Rieble, C., Fried, E. I., & Liu, S. (2024). *Assessing the internal consistency reliability of ecological momentary assessment measures: Insights from the WARN-D study*. OSF. https://doi.org/10.31234/osf.io/nrzsc

Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. *Psychological Assessment*, *26*(2), 513–527. https://doi.org/10.1037/a0035768

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917–929. https://doi.org/10.1177/0146167206287721

Crowe, E., Daly, M., Delaney, L., Carroll, S., & Malone, K. M. (2019). The intra-day dynamics of affect, self-esteem, tiredness, and suicidality in Major Depression. *Psychiatry Research*, *279*, 98–108. https://doi.org/10.1016/j.psychres.2018.02.032

Degroote, L., DeSmet, A., De Bourdeaudhuij, I., Van Dyck, D., & Crombez, G. (2020). Content validity and methodological considerations in ecological momentary assessment studies on physical activity and sedentary behaviour: A systematic review. *The International Journal of Behavioral Nutrition and Physical Activity*, *17*(1), Article 35. https://doi.org/10.1186/s12966-020-00932-9

Dubad, M., Winsper, C., Meyer, C., Livanou, M., & Marwaha, S. (2018). A systematic review of the psychometric properties, usability and clinical impacts of mobile mood-monitoring applications in young people. *Psychological Medicine*, *48*(2), 208–228. https://doi.org/10.1017/S0033291717001659

Ebner-Priemer, U. W., & Trull, T. J. (2009). Ecological momentary assessment of mood disorders and mood dysregulation. *Psychological Assessment*, *21*(4), 463–475. https://doi.org/10.1037/a0017075

Fisher, R. A. (1921). On the" Probable Error" of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32.

Haddox, D. (2025). *Mobile patient health questionnaire-9 (MPHQ-9)*. https://osf.io/g5fx3

Haslbeck, J., Jover Martínez, A., Roefs, A., Fried, E., Lemmens, L. H. J. M., Groot, E. L., & Edelsbrunner, P. A. (2024). Comparing likert and visual analogue scales in ecological momentary assessment. *Behavior Research Methods*, *57*(8), Article 217. https://doi.org/10.31234/osf.io/yt8xw

Haslbeck, J., Ryan, O., & Dablander, F. (2023). Multimodality and skewness in emotion time series. *Emotion*, *23*(8), 2117–2141. https://doi.org/10.1037/emo0001218

Horwitz, A. G., McCarthy, K., & Sen, S. (2024). A review of the peak-end rule in mental health contexts. *Current Opinion in Psychology*, *58*, Article 101845. https://doi.org/10.1016/j.copsyc.2024.101845

Horwitz, A. G., Zhao, Z., & Sen, S. (2023). Peak-end bias in retrospective recall of depressive symptoms on the PHQ-9. *Psychological Assessment*, *35*(4), 378–381. https://doi.org/10.1037/pas0001219

Jamalabadi, H., Stocker, J. E., Koosha, T. A., Jansen, A., Ebner-Priemer, U., Rieble, C. L., Proppert, R. K. K., Tutunji, R., & Fried, E. I. (2024). *Optimizing the frequency of ecological momentary assessments using signal processing*. https://doi.org/10.31234/osf.io/hev28

Jimenez, A., McMahon, T. P., Watson, D., & Naragon-Gainey, K. (2022). Dysphoria and well-being in daily life: Development and validation of ecological momentary assessment scales. *Psychological Assessment*, *34*(6), 546–557. https://doi.org/10.1037/pas0001117

Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H. A., Wen, C. K. F., & Field, M. (2019). Compliance with ecological momentary assessment protocols in substance users: A meta-analysis. *Addiction*, *114*(4), 609–619. https://doi.org/10.1111/add.14503

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*(6), 401–405. https://doi.org/10.1111/j.1467-9280.1993.tb00589.x

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Liu, M., & Conrad, F. G. (2019). Where should I start? On default values for slider questions in web surveys. *Social Science Computer Review*, *37*(2), 248–269. https://doi.org/10.1177/0894439318755336

Neubauer, A. B., & Schmiedek, F. (2020). Studying within-person variation and within-person couplings in intensive longitudinal data: Lessons learned and to be learned. *Gerontology*, *66*(4), 332–339. https://doi.org/10.1159/000507993

Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality*, *69*, 149–155. https://doi.org/10.1016/j.jrp.2016.06.020

Redelmeier, D. A., Katz, J., & Kahneman, D. (2003). Memories of colonoscopy: A randomized trial. *Pain*, *104*(1), 187–194. https://doi.org/10.1016/S0304-3959(03)00003-4

Revelle, W. (2025). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.5.3) [R package]. Northwestern University. https://CRAN.R-project.org/package=psych

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). McGraw-Hill.

Rush, A. J. (2015). Isn't it about time to employ measurement-based care in practice? *The American Journal of Psychiatry*, *172*(10), 934–936. https://doi.org/10.1176/appi.ajp.2015.15070928

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*(3), 301–317. https://doi.org/10.1177/096228029800700306

Stone, A. A., Schneider, S., & Smyth, J. M. (2023). Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology*, *19*(1), 107–131. https://doi.org/10.1146/annurev-clinpsy-080921-083128

Targum, S. D., Sauder, C., Evans, M., Saber, J. N., & Harvey, P. D. (2021). Ecological momentary assessment as a measurement tool in depression trials. *Journal of Psychiatric Research*, *136*, 256–264. https://doi.org/10.1016/j.jpsychires.2021.02.012

Torous, J., Staples, P., Shanahan, M., Lin, C., Peck, P., Keshavan, M., & Onnela, J.-P. (2015). Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. *JMIR Mental Health*, *2*(1), Article e8. https://doi.org/10.2196/mental.3889

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 3–14). https://doi.org/10.1145/2632048.2632054

Wang, S., Yang, C.-H., Brown, D., Cheng, A., & Kwan, M. Y. W. (2025). Participant compliance with ecological momentary assessment in movement behavior research among adolescents and emerging adults: Systematic review. *JMIR MHealth and UHealth*, *13*, Article e52887. https://doi.org/10.2196/52887

Watson, D., O'Hara, M. W., Naragon-Gainey, K., Koffel, E., Chmielewski, M., Kotov, R., Stasik, S. M., & Ruggero, C. J. (2012). Development and validation of new anxiety and bipolar symptom scales for an expanded version of the IDAS (the IDAS-II). *Assessment*, *19*(4), 399–420. https://doi.org/10.1177/1073191112449857

Wichers, M., Riese, H., Hodges, T. M., Snippe, E., & Bos, F. M. (2021). A narrative review of network studies in depression: What different methodological approaches tell us about depression. *Frontiers in Psychiatry*, *12*, Article 719490. https://doi.org/10.3389/fpsyt.2021.719490

Winkielman, P., Knäuper, B., & Schwarz, N. (1998). Looking back at anger: Reference periods change the interpretation of emotion frequency questions. *Journal of Personality and Social Psychology*, *75*(3), 719–728. https://doi.org/10.1037/0022-3514.75.3.719

Wirz-Justice, A. (2008). Diurnal variation of depressive symptoms. *Dialogues in Clinical Neuroscience*, *10*(3), 337–343. https://doi.org/10.31887/DCNS.2008.10.3/awjustice

World Health Organization. (2017). *Depression and other common mental disorders: Global health estimates* (WHO/MSD/MER/2017.2). Article WHO/MSD/MER/2017.2. https://iris.who.int/handle/10665/254610

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, *30*(3), 825–846. https://doi.org/10.1177/10731911211067538

Received November 7, 2024
Revision received September 3, 2025
Accepted September 5, 2025 ■