JMIR Publications
The Leading eHealth Publisher

Are ecological momentary assessment measures of intervention change worth the trouble? Evaluation in four digital mental health trials

Christian Webb, Lori Hilt, Caroline Swords, Daniel Bolt, Hadar Fisher, Simon Goldberg

Submitted to: Journal of Medical Internet Research
on: November 26, 2024

Preprint
JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Preprint
JMIR Publications

Are ecological momentary assessment measures of intervention change worth the trouble? Evaluation in four digital mental health trials

Christian Webb[1, 2] PhD; Lori Hilt[3] PhD; Caroline Swords[3] BA; Daniel Bolt[4] PhD; Hadar Fisher[2, 1] PhD; Simon Goldberg[4] PhD

[1]McLean Hospital Belmont US
[2]Harvard Medical School Boston US
[3]Lawrence University Appleton US
[4]University of Wisconsin–Madison Madison US

Corresponding Author:
Christian Webb PhD
McLean Hospital
115 Mill St
Belmont
US

Abstract

Ecological momentary assessment (EMA) is increasingly being incorporated into intervention studies to acquire a more fine-grained and ecologically valid assessment of change. The added utility of including relatively burdensome EMA measures in a clinical trial hinges on several psychometric assumptions, including that these measure are: (1) reliable, (2) related but not redundant with conventional self-report measures (convergent and discriminant validity), (3) sensitive to intervention-related change, and (4) associated with a clinically-relevant criterion of improvement (criterion validity) above conventional self-report measures (incremental validity). Using data from 4 app-based meditation trials (N = 412), we examined the reliability, validity, and sensitivity to change of conventional self-report and EMA measures of improvement in rumination. Conventional self-report and EMA measures of rumination were only modestly correlated, particularly with regards to change over time, which may be due to their lower reliability. Changes in rumination were larger for conventional self-report than EMA. Notably, change in both self-report and EMA rumination each accounted for unique variance in depressive symptom improvement, demonstrating incremental predictive validity. Conventional self-report and EMA measures of rumination provide distinct and clinically meaningful information. Researchers using EMA should consider the psychometric properties of their measures and the precise construct they intend to capture.

(JMIR Preprints 26/11/2024:69297)
DOI: https://doi.org/10.2196/preprints.69297

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Preprint JMIR Publications

Original Submission: *November 26, 2024*

Are ecological momentary assessment measures of intervention change worth the trouble? Evaluation in four digital mental health trials

Christian A. Webb[1,2], Lori M. Hilt[3], Caroline M. Swords[3,5],

Daniel M. Bolt [4], Hadar Fisher[1,2], and Simon B. Goldberg[5,6]

[1]Harvard Medical School, Department of Psychiatry, Boston, MA

[2]McLean Hospital, Center for Depression, Anxiety & Stress Research, Belmont, MA

[3]Lawrence University, Department of Psychology, Appleton, WI

[4]University of Wisconsin, Madison, Department of Educational Psychology, Madison WI

[5]Center for Healthy Minds, University of Wisconsin – Madison, Madison, WI

[6]Department of Counseling Psychology, University of Wisconsin – Madison, Madison, WI

Corresponding author:
Christian A. Webb, Ph.D.
Harvard Medical School & McLean Hospital
115 Mill Street, Belmont, MA 02478
Email: cwebb@mclean.harvard.edu

Preprint JMIR Publications

Abstract

Background: Ecological momentary assessment (EMA) is increasingly being incorporated into intervention studies to acquire a more fine-grained and ecologically valid assessment of change. The added utility of including relatively burdensome EMA measures in a clinical trial hinges on several psychometric assumptions, including that these measure are: (1) reliable, (2) related but not redundant with conventional self-report measures (convergent and discriminant validity), (3) sensitive to intervention-related change, and (4) associated with a clinically-relevant criterion of improvement (criterion validity) above conventional self-report measures (incremental validity).

Method: Using data from 4 app-based meditation trials (N = 412), we examined the reliability, validity, and sensitivity to change of conventional self-report and EMA measures of improvement in rumination.

Results: Conventional self-report and EMA measures of rumination were only modestly correlated, particularly with regards to change over time, which may be due to their lower reliability. Changes in rumination were larger for conventional self-report than EMA. Notably, change in both self-report and EMA rumination each accounted for unique variance in depressive symptom improvement, demonstrating incremental predictive validity.

Conclusions: Conventional self-report and EMA measures of rumination provide distinct and clinically meaningful information. Researchers using EMA should consider the psychometric properties of their measures and the precise construct they intend to capture.

Keywords: ecological momentary assessment; rumination; mindfulness; apps; reliability

Preprint JMIR Publications

Introduction

An increasing number of intervention studies (particularly mobile mental health trials) are incorporating smartphone-delivered ecological momentary assessment (EMA) to measure symptom improvement (Hilt & Swords, 2021; Mofsen et al., 2019; Moore et al., 2016; Peterson et al., 2020; Targum et al., 2021; Webb, Swords, et al., 2021; Webb et al., 2022, 2023; Wichers et al., 2009). Proponents of EMA argue that the reliance on commonly used retrospective self-report questionnaires may be convenient and cost-effective, but these measures often lack ecological validity and are contaminated by recall bias (Mofsen et al., 2019; Stone & Shiffman, 2010; Van den Bergh & Walentynowicz, 2016). Specifically, in contrast to EMA surveys which are deployed in daily life and ask about current or recent (e.g., "since the last survey") experiences, conventional self-report symptom questionnaires typically ask participants to recall and summarize their average levels of symptoms over a relatively lengthy period (e.g., several weeks). For example, the commonly used Beck Depression Inventory (Beck et al., 1996) and Center for Epidemiologic Studies Depression Scale (Andresen et al., 1994) ask participants to recall and summarize their experience of depressive symptoms over the past 2 weeks and 1 week, respectively.[1] At an even grosser level of temporal abstraction, trait or global self-report measures ask respondents to report on their *typical* tendencies or personal characteristics without specifying a recall timeframe. For example, the Response Styles Questionnaire (RSQ) (Nolen-Hoeksema, 1991) and Children's Response Styles Questionnaire (CRSQ) (Abela et al., 2002) are commonly used to assess depressive rumination in adults and children, respectively. These measures ask participants what they "generally" or "usually" do when feeling sad.

Prior research reveals that retrospective reports are often biased. For example, studies show that both children (Chen et al., 2000; Noel et al., 2015) and adults (Broderick et al., 2008; Friedberg

[1] It is worth noting that common clinician-administered symptom measures (e.g., Hamilton Depression Rating Scale (Hamilton, 1960) or Children's Depression Rating Scale (Poznanski & Mokros, 1996)) are not immune from these issues as they are still based on participant self-report to the interviewer and ask about timeframes which extend into the past. However, skillful clinician questioning may improve participant recall and reduce memory biases relative to conventional self-report questionnaires.

Preprint JMIR Publications

& Sohl, 2008; Giske et al., 2010; Rinner et al., 2019; Stone et al., 2004) overestimate levels of symptoms on retrospective self-report measures relative to averaged EMA reports (for evidence of age moderating this effect, see Neubauer et al., 2020 & Zurbriggen et al., 2021). Studies have also shown that when individuals report on their experiences (e.g., pain during a medical procedure) via retrospective self-report they are often incorporating information about the most intense (peak) moment and how the experience ended (i.e., the so called “peak-end” bias), rather than equally weighting timepoints and simply averaging their momentary experiences over the entire reporting timeframe (Chajut et al., 2014; Kahneman et al., 1993; Redelmeier et al., 2003; Redelmeier & Kahneman, 1996). This “memory-experience gap” is due to the fact that whereas momentary EMA requires introspection of current states (e.g., emotions, thoughts and behaviors in the “here-and-now”), retrospective self-report taps autobiographical (episodic) memory, which is limited by the extent to which past experiences and events have been accurately encoded and consolidated in memory (Conner & Barrett, 2012; Miron-Shatz et al., 2009; Van den Bergh & Walentynowicz, 2016). Moreover, memory retrieval has been shown to be biased by one’s current state (e.g., mood-congruent memory bias; Faul & LaBar, 2022). This is further complicated by the fact that individuals may shift from episodic memory to semantic memory (or a combination of the two) when recall timeframes are longer or remembering relevant details of past episodes becomes challenging (Robinson & Clore, 2002; Van den Bergh & Walentynowicz, 2016). Semantic knowledge includes more abstract beliefs or generalization about the self which are not tied to a specific time and place (e.g., “I’m generally a happy person,” “I’m a worrier”; Robinson & Clore, 2002; Tulving, 2002; Van den Bergh & Walentynowicz, 2016). One study indicated that the shift from relying on episodic memory in reporting past emotional well-being to drawing on both semantic and episodic retrieval strategies was in the 3 to 7 week range (Geng et al., 2013) (but see Walentynowicz et al., 2018)

Collectively, these findings suggest that intervention researchers may benefit from relying less on convenient but often biased retrospective self-report measures and instead adopt EMA to

Preprint JMIR Publications

measure symptom improvement. However, EMA does not necessarily provide an inherently more reliable and valid measure of intervention-related change.[2] The repeated assessments involved in EMA protocols may introduce unique biases and careless responding (e.g., random or stereotyped responding) among at least some participants (Jaso et al., 2022). For example, there is some evidence of an "initial elevation bias" (i.e., an upward bias in scores on initial self-reports) when subjective reports are collected repeatedly, and this effect is more pronounced for negative mental states and physical symptoms than for positive states and behaviors (Shrout et al., 2018). However, a subsequent study found these effects to be inconsistent and minimal (Cerino et al., 2022). As another example, a recent study applied drift diffusion modeling to reaction time (RT) data from EMA affect ratings and found evidence of a shift over time in the cognitive processes that underlie survey responses. Specifically, the authors found pronounced changes in two central drift diffusion parameters across the repeated EMA affect ratings: increasing "drift rate" (reflecting faster processing of affective information over time) and decreased "boundary separation" (reflecting a decrease in how thoroughly participants process items and how cautiously they respond over time) (Schneider et al., 2024; but see Hernandez et al., 2024). In summary, at this stage, given conflicting findings, it is not entirely clear to what extent EMAs are more reliable and valid measures relative to conventional retrospective self-report instruments.

The added utility of including relatively burdensome EMA measures of change in an intervention study hinges on several psychometric assumptions, including that these measure are: (1) reliable, (2) related but not redundant with conventional self-report measures of the same construct (convergent and discriminant validity), (3) sensitive to intervention-related change, and (4) associated with a clinically-relevant criterion of patient improvement (criterion validity) above and beyond conventional self-report measures (incremental validity). In the present study, we include data from four clinical trials (*N* = 412) of app-delivered meditation training which incorporated both

[2] We focus here on EMA measures of improvement in rumination levels over time. It is important to note that EMA can be a useful tool for other clinically relevant questions, such as investigating dynamics (e.g., in emotions or symptoms), within-person variability, lagged associations over short-time frames and contextual associations.

Preprint JMIR Publications

EMA and conventional self-report measures of improvement in rumination. We compared these measures with regards to their reliability, validity, and sensitivity to detecting change in order to help inform which measures to include in future intervention research.

Methods

Participants

There were two adolescent samples, aged 12-15, selected for moderate-to-high levels of rumination from the community. The 80 participants in Sample 1 (*M* age = 14.01, *SD* = 0.99; 45.0% girls; 86.3% White; 3.8% Hispanic; Median income: $100,000–125,000) were recruited from a midsized midwestern city in the USA in 2018-2019 for a single arm trial investigating a mindfulness mobile application. Ninety percent (70/80) of the participants completed the 3-week trial (for more details, see Hilt & Swords, 2021). The 152 participants in Sample 2 *(M* age = 13.71 years, *SD* = 0.89; 58.6% girls; 82.2% White; 10.5% Hispanic; Median income: $90,000–100,000) were recruited in 2019-2020 from the same area for a randomized controlled trial investigating a mindfulness mobile application. Eighty-nine percent (136/152) of the participants completed the trial (for more details, see Hilt et al., 2023).

Sample 3 included 88 first-year undergraduate students (recruited from the same area as Samples 1 and 2) aged 18-21 (*M* age = 18.51, *SD* = 0.64; 65.9% female, 67% White, 11.4% Hispanic; income data not collected) recruited in 2018 for a randomized controlled trial investigating a mindfulness mobile application. Ninety percent (79/88) of the participants completed the trial (for more details, see Hilt & Swords, 2021).

Sample 4 included 92 adults from the Madison, Wisconsin area with elevated levels of depression and/or anxiety (i.e., Patient-Reported Outcomes Information System [PROMIS] Depression and/or PROMIS Anxiety T-scores > 55; Choi et al., 2014; Schalet et al., 2014) ages 18 to 80 (*M* age = 31.28, *SD* = 13.12; 81.5% female, 69.6% White, 8.7% Hispanic; Median income $45,000) recruited in 2021 for a randomized trial investigating a meditation-based mobile

Preprint JMIR Publications

application. Ninety-seven percent (89/92) completed the trial (for more details, see https://classic.clinicaltrials.gov/ct2/show/NCT05229406).

Power analyses for Sample 1 (Hilt & Swords, 2021), Sample 2 (Hilt et al., 2023) and Sample 3 (Hilt, Vachhani, Swords, Vaghasia, & Sage, under review; and described in Swords & Hilt, 2021 where it was used for another purpose) have been previously reported. Sample 4 (unpublished) was designed to test the feasibility and acceptability of manipulating practice dosage. It was not powered to detect a particular effect. See preregistration https://osf.io/fszvj/?view_only=9cb1b9e67cc042f9bc7a0309e94b2f52 and https://classic.clinicaltrials.gov/ct2/show/NCT05229406.

Procedure

Adolescent participants (Samples 1 and 2) were recruited through letters sent to parents in the local school district and by word of mouth. They were invited to participate if a two-item phone screen for rumination determined that they reported responding to sadness or stress with rumination at least "sometimes." At baseline, adolescents and parents completed self-report questionnaires, set up the mindfulness mobile application (Child and Adolescent Research in Emotion [CARE] app) on the adolescent's mobile device, and learned how to use it. In Sample 1, all participants used the same version of the app, which prompted participants to report on mood and rumination and included a mindfulness intervention. In Sample 2, adolescents were randomly assigned to use either a mood monitoring control version of the app or the mindfulness condition of the app that Sample 1 used (see below). For three weeks, adolescents in both samples were prompted to use the CARE app three times a day by notifications sent through the application. After the intervention period, participants completed online follow-up questionnaires that included the same measure of trait rumination completed as part of their baseline questionnaires.

College students (Sample 3) were recruited from introductory classes, at events for first-year students, and with fliers. Participants were eligible to participate if they did not report serious

Preprint JMIR Publications

suicidal concerns. Participants completed baseline questionnaires that included a measure of trait rumination and were randomized to either the mindfulness version of the CARE app or the mood monitoring control condition at baseline. During the three-week intervention period, participants were prompted to use the app three times a day by notifications sent to their phone. At post-intervention, participants completed questionnaires that included the same measure of trait rumination completed at baseline.

Adults with elevated depression and/or anxiety (Sample 4) were recruited through flyers placed in the community and recruitment emails sent to faculty, staff, and students at the University of Wisconsin – Madison. Participants were eligible if they were 18 years or older; had access to a smartphone capable of downloading the Healthy Minds Program (HMP) app; self-reported willingness to complete EMA for 4 weeks; were able to speak, read, and write in English; and had clinically elevated PROMIS Depression and/or Anxiety (T-scores > 55). Exclusion criteria included: prior meditation retreat experience; a regular meditation practice (weekly practice for >1 year or daily practice within the previous 6 months); previous practice under the instruction of a meditation teacher; severe depression symptoms (PROMIS Depression > 70); or a positive screen for alcohol dependence on the Alcohol Use Disorders Identification Test (AUDIT; Aalto et al., 2009).

Conditions

Mood Monitoring Control Condition

Samples 2 and 3 used a mood monitoring control condition through the CARE app. Participants assigned to the mood monitoring control condition were prompted to report on state mood and rumination three times a day without the chance of receiving a mindfulness intervention. After receiving a notification to use the app, participants reported on their current mood (i.e., sad, anxious, happy, and calm) and state rumination on a scale of 0 (*not at all*) to 100 (*extremely*). For more information, see Hilt, Swords & Webb, 2023.

Meditation Condition

Preprint JMIR Publications

Samples 1, 2, and 3 used a mindfulness meditation condition through the CARE app. Participants assigned to the mindfulness condition answered the same mood and rumination questions from the mood monitoring control condition and they also received mindfulness exercises. To prevent participants from learning which responses would result in receiving a mindfulness exercise, there was a 67% probability of receiving an exercise each time the app was used. If participants indicated high levels of anxiety or sadness (i.e., ≥ 90), their chances of receiving a mindfulness exercise increased to 85%. If prompted to receive a mindfulness exercise, participants could select how much time they had to complete an exercise (i.e., about 1, 5, or 10 minutes), and an exercise was randomly assigned within those parameters. After completing the mindfulness exercise, participants reported again on their current mood and rumination. For more details, see Hilt, Swords & Webb, 2023.

Sample 4 used the 4-week Foundations program from the HMP mobile app. HMP includes meditation practices linked with four dimensions of well-being: awareness, connection, insight, and purpose (ACIP; Dahl et al., 2020). Briefly, awareness practices aim to cultivate mindfulness and attention regulation, connection practices aim to cultivate healthy relationships with oneself and others, insight practices aim to cultivate an understanding of how our internal experience (e.g., emotions, thoughts) shape our experience, and purpose practices aim to cultivate a connection with one’s values and a sense of meaning in daily life. For more details, see Goldberg et al. (2020) and Hirshberg et al. (2022). Within this study, participants were randomly assigned to use HMP for either 5- or 15-minutes per day (defined as low- and high-dose conditions). In addition to using HMP, participants were sent EMAs assessing various aspects of well-being (ACIP dimensions), psychological distress (depression and anxiety), stressor exposure, and rumination 4 times per day over the 4-week intervention period.

Measures

Self-report depressive symptoms

Preprint JMIR Publications

For adolescents (Samples 1 and 2), depressive symptoms were assessed with the Children's Depression Inventory (CDI; Kovacs, 1992). The CDI is a 27-item measure of depression adapted for children and adolescents from the Beck Depression Inventory (BDI; Beck et al., 1996). The CDI assesses the frequency and severity of depressive symptoms over the past two weeks. Participants report on symptoms and severity of depressive symptoms using a 4-point scale with higher scores on the CDI indicating greater frequency and severity of symptoms. Past research demonstrates that the CDI is reliable and valid in child and adolescent samples (Craighead et al., 1995; Klein et al., 2005). In Sample 1, the CDI showed good reliability at baseline (α = .82), post-intervention (α = .86), 6-week follow-up (α = .88) and 12-week follow-up (α = .91). In Sample 2, the CDI showed excellent reliability at baseline (α = .90), post-intervention (α = .92), 6-week follow-up (α = .91) and 12-week follow-up (α = .92).

The Beck Depression Inventory (BDI-II; Beck et al., 1996) was used to assess depressive symptoms in college students (Sample 3). The BDI-II is a 21-item measure that asks participants to self-report on depressive symptoms experienced in the past two weeks using a 4-point scale. Higher scores on the BDI-II indicate greater symptom frequency and severity. The BDI-II has demonstrated validity and reliability in college samples (Storch et al., 2004). The measure demonstrated excellent reliability at baseline (α = .94), post-intervention (α = .93), 6-week follow-up (α = .88) and 12-week follow-up (α = .91).

The computer-adaptive version of the PROMIS Depression scale (Pilkonis et al., 2011) was used to assess depressive symptoms in adults with elevated depression and/or anxiety (Sample 4). The computer-adaptive PROMIS Depression draws items from a bank of 28 items. Participants rate their experience of various depression symptoms in the past 7 days on a 5-point scale ranging from 1 (*never*) to 5 (*always*). The measure has shown strong convergent validity with legacy measures of depression including the BDI-II (Choi et al., 2014). As not all participants receive the same items, internal consistency cannot be calculated. However, short-form versions of the PROMIS Depression

Preprint JMIR Publications

scale have shown excellent reliability (e.g., α = .97 for the 8-item PROMIS Depression; Nolte et al., 2019).

Self-report trait rumination

Adolescent self-report trait rumination (Samples 1 and 2) was assessed with the 13-item rumination subscale of the Children's Response Styles Questionnaire (CRSQ; Abela et al., 2002). For each item, participants are instructed to report on how they respond to feelings of sadness *or stress*. Instructions were modified to include stress, in line with current conceptualizations of rumination (Nolen-Hoeksema et al., 2008). In response to each item, participants indicate whether they respond in the way described by the item on a scale from 0 (*almost never)* to 4 (*almost always*). Higher scores indicate greater rumination. Past research suggests that the CRSQ is reliable and valid in an adolescent sample (Abela et al., 2002). In the present study, the rumination subscale of the CRSQ demonstrated good reliability in sample 1 (α = .89) and excellent reliability in sample 2 (α = .92) at baseline. At post-intervention, the CRSQ showed excellent reliability in sample 1 (α = .90) sample 2 (α = .92).

College student rumination (Sample 3) was assessed with the ruminative response subscale (RRS) from the Response Styles Questionnaire (RSQ; Nolen-Hoeksema & Morrow, 1991). The RRS is a 22-item measure in which participants are asked to rate how often they respond as described by the item on a scale from 0 (*almost never*) to 4 (*almost always*). Higher scores indicate greater frequency and severity of rumination. Past research in college student samples suggests that the RRS has good internal consistency and moderate test-retest reliability (Roelofs et al., 2006). Similar to the CRSQ above, instructions were adapted to ask participants to report on their response to sadness *or stress* (Nolen-Hoeksema et al., 2008). In this sample, the RRS showed excellent reliability at baseline (α = .93) and at follow up (α = .94).

Repetitive negative thinking in the adult sample with elevated depression and/or anxiety (Sample 4) was assessed using the Perseverative Thinking Questionnaire (PTQ; Ehring et al., 2011).

Preprint JMIR Publications

It should be noted that repetitive negative thinking is a broader construct than rumination (not solely focused on past-oriented repetitive negative thoughts but also includes, for example, future-oriented worry and present concerns). The PTQ is a 15-item measure where participants are asked how often they respond in a particular way on a scale from 0 (*never*) to 4 (*almost always*). Items query various forms of repetitive negative thinking (e.g., “I keep thinking about the same issue all the time”). Higher scores indicate greater frequency of repetitive negative thought. Internal consistency and test-retest reliability along with convergent and predictive validity have been established for adults (Ehring et al., 2011). In this sample, the PTQ showed excellent reliability at baseline (α = .94) and at follow up (α = .95).

EMA rumination

In Samples 1, 2, and 3, state rumination was assessed by asking participants “How much were you focusing on your emotions?” and “How much were you focusing on your problems?” “just before” seeing the prompt to use the app. Participants rated the degree to which they had been ruminating, as described by these questions, on a scale of 0 (*not at all*) to 100 (*extremely*). These questions were created in line with past research (Hilt & Pollak, 2012; Moberly & Watkins, 2008). In the analyses below, we focus on the latter question (problem-focused rumination) to assess rumination given (1) some concerns about the extent to which focusing on emotions in fact assesses problematic rumination (e.g., participants may direct *greater* attention to their emotions by virtue of mindfulness training) (Webb et al., 2022) (also see Hilt et al., 2017, which focused on problem-focused rumination), (2) the relatively modest correlation between problem-focused and emotion-focused rumination (r = 0.41 in this sample, which raises questions regarding internal consistency, (3) and the fact that these two items yield different patterns of findings in past work (Webb et al., 2022). Participants in Samples 1-3 completed a mean of 74.1%, 79.4% and 72.2% surveys during the 3-week intervention period, respectively.

In Sample 4, participants first completed a stressor exposure item which asked participants to

Preprint JMIR Publications

"Think about the most stressful or negative thing that happened since you completed the last survey" and then to indicate how stressed they felt "at the worst point" on a scale from 1 (*not at all*) to 7 (*very much*). Rumination was then assessed with a subsequent item: "After the stressful or negative thing happened, I was dwelling on my mistakes, failures, or losses" which was also rated on a scale from 1 (*not at all*) to 7 (*very much*). These items were also created in line with past research (Ruscio et al., 2015; Webb, Israel, et al., 2021). EMA compliance was 81.7%.

Analysis code and Sample 4 data are available on OSF: https://osf.io/2bnmk/. Data for Samples 1-3 are available upon request (IRB requires a signed data sharing agreement). The data analytic plan for this study was not preregistered. Original trial registration for Sample 2 can be found at clinicaltrials.gov (NCT03900416). Original trial registration for Sample 4 can also be found at OSF (https://osf.io/fszvj/?view_only=9cb1b9e67cc042f9bc7a0309e94b2f52) and clinicaltrials.gov (NCT05229406). We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. All participants completed informed consent procedures approved by that Institutional Review Board at Lawrence University (Samples 1-3) or the University of Wisconsin (Sample 4).

Analytic Strategy

Reliability of self-report and EMA measures

Prior to estimating the reliability of rumination change scores, we first report the reliability (internal consistency) of conventional retrospective self-report rumination (for brevity, we refer to the latter as simply "retrospective rumination" below) at baseline (T1) and post-intervention (T2). Next, to compute the reliability of residualized change in retrospective rumination we use the formula below which has the following variables as inputs: the reliability of retrospective rumination scores at T1 (α_{T1}¿ and T2 (α_{T2}¿, and the squared correlation coefficient ($\rho^2_{T1,T2}$¿ between these two scores (Z; Williams & Zimmerman, 1982 Eq 2):

Preprint JMIR Publications

$$Reliability(Z)=\frac{\alpha_{T2}+\rho_{T1,T2}^{2}\alpha_{T1}-2\rho_{T1,T2}^{2}}{1-\rho_{T1,T2}^{2}}$$

Note that the reliability of residualized change (Z) will increase if the reliability of either T1 or T2 rumination scores increases or if the correlation between T1 and T2 rumination *decreases* (see Supplement for a formula for the reliability of raw, rather than residualized, change). Next, to approximate the reliability of EMA rumination we computed a type of split-half reliability estimate evaluating the reliability of the participant mean EMA score over time (National Center on Intensive Intervention, 2014; Van Norman & Parker, 2018). Specifically, we *randomly* split each participant's vector of EMA observations into two subvectors, computed mean rumination in each subvector for each participant and correlated these scores. To approximate reliability of change over time, we computed the ordinary least squares (OLS) linear slope for rumination within each subvector (i.e., for each participant, computed the slope in the two random subsets of EMA observations) and correlated these slopes.[3] An alternative approach to evaluating the reliability of OLS slopes estimates this reliability *within* a multilevel model (MLM) that includes individual-level random intercepts and slopes (see Raudenbush & Bryk, 2002; pp. 49-50 for details). Some evidence suggests reliability estimates are slightly higher for this second approach (VanNorman & Parker, 2018). We applied both the split-half and MLM methods to observe this distinction, although consistently report the former split-half approach both to provide greater consistency to the approach taken with self-report measures and because of its frequent use in practice.

Given that the measures of reliability for retrospective rumination (internal consistency) and EMA rumination (split-half reliability) remain somewhat different, they are not directly comparable. An important aspect of their distinction is that each is attending to different sources of error variability in how the reliability of change is characterized. As noted above, our self-report reliability

[3] Another method for evaluating change in this context would attend to individual-level Empirical Bayes' estimates of the slopes. Such estimates can be viewed as OLS estimates adjusted for lack of reliability. This approach was not taken here due to the tendency for such estimates to be "shrunken", thus potentially introducing bias into analyses evaluating the relationships between slopes and other variables (either predictor of the slopes or outcomes) (Liu et al., 2021) (see Supplement).

estimate attends to item-related error, which is ignored (or irrelevant) in the EMA approach due to the availability of only one item; similarly, time-related error is ignored (or irrelevant) in the self-report due to a definition of change that incorporates time-related error into true change. In analyses presented in the supplement, we demonstrate (using one of our EMA datasets in which a second item could be included) how a data collection design that allows quantification of both sources of error allows assessment of various reliability coefficients that could applied to estimate reliability coefficients relevant to different conditions. Unfortunately, none of these would be applicable to our evaluation of the reliability of the OLS slope so are not included in our main analyses. We can, however, evaluate the decrement (if any) in reliability within each measure when examining change over time.

Relationship between conventional self-report and EMA measures

Preprint JMIR Publications

Similar to above, before testing the association between change scores, we first tested the correlation between baseline (T1) scores on the retrospective rumination measure and mean EMA rumination over the initial 3 days of EMA. Similarly, we correlated retrospective rumination at post-intervention (T2) and mean EMA rumination over the prior 3 days. We selected a 3-day window for averaging in an effort to obtain a relatively representative estimate of an individual's typical tendency to ruminate (see Supplement for analyses averaging over different numbers of days, which yielded similar findings). We also tested the association between change in retrospective rumination and EMA rumination from T1 to T2. To compute change in retrospective rumination, we saved the residuals from a model in which T2 retrospective rumination scores served as the dependent variable and T1 retrospective rumination was the predictor variable, creating a residualized change score. To compute change in EMA rumination, the slope of rumination change over the course of the trial (i.e., 3 weeks / 21 days for Samples 1, 2, and 3; 4 weeks / 28 days for Sample 4) was computed from subject-specific regressions[4] of rumination scores on intervention day (National Center on Intensive Intervention, 2014; Van Norman & Parker, 2018). Given the differences between Samples 1-3 (which all used the identical CARE mindfulness app for 3 weeks in samples of adolescents and college students) and Sample 4 (used the HMP meditation app for 4 weeks in adults) all analyses were conducted separately for the CARE (with sample included as a covariate in analyses) and HMP studies (henceforth referred to as the CARE and HMP samples, respectively).

Comparing rumination improvement from self-report and EMA measures

Given that the CARE sample included a mood monitoring control condition, we tested whether group differences emerged in change in rumination from T1 to T2. For retrospective rumination, we ran a regression predicting T2 rumination (adjusting for T1 rumination scores) with Group as the predictor of interest. For EMA rumination, a robust linear mixed effects models

[4] As our focus in this analysis is solely on obtaining and interpreting participant-level estimates of systematic linear change over time, alternative multilevel reliability approaches are perceived as less relevant (but see the supplement on multilevel reliability, which illustrates the application of such methods).

(LMM), using the statistical package ‘robustlmm’ (version 3.1) in R (Version 4.2.2), with a Group x *Time* interaction was specified (with random intercepts and slopes). A standardized effect size for the Group x *Time* interaction was estimated using the $d_{GMA\text{-}RAW}$ (β_{11}(time)/SD_{RAW}) formula recommended by Feingold (2009) (for brevity $d_{GMA\text{-}RAW}$ is referred to as simply *d* below).

For the HMP sample, in which both groups received the HMP app, we examined whether EMA rumination changed over time. To do this, we fit a robust LMM with EMA rumination as the dependent variable and *Time* in days (ranging from 0 to 28) as the independent variable, with the nesting of observations within participants modeled using a random intercept and slope. A standardized effect size for the effect of *Time* was estimated following the above $d_{GMA\text{-}RAW.}$formula.

Criterion validity of self-report and EMA measures

Finally, to test whether change in traditional self-report rumination and/or EMA rumination predicts improvement in depression severity (criterion), we used robust LMMs. Consistent with prior work arguing that targeting repetitive negative thinking (rumination) improves depressive symptoms (Hilt et al., 2017; Webb et al., 2022; Webb, Swords, et al., 2021), these analyses are an attempt to test if improvement in either retrospective rumination or EMA rumination is related to improvement in depression (NCT03900416, NCT06348277 and NCT05229406). For the CARE sample, the dependent variable incorporated repeated assessment of depression over 4 timepoints (i.e., baseline, post-treatment, 6 weeks follow-up, and 12 weeks follow-up) which were nested within individuals. *Time* was centered to represent estimated symptom scores at the final follow-up (12 weeks) and baseline depression was included as a covariate. Random intercepts and slopes were specified. To test whether either change (baseline to post-treatment) measure predicted improvement in depression over time (baseline through 12-week follow-up), we included each term in an interaction with *Time* (i.e., retrospective rumination change x *Time* and EMA rumination change x *Time*). Each term was simultaneously included in the same model.

For the HMP sample, traditional self-reported depression was assessed only at pre- and post-test. We therefore used ordinary least squares regression models with post-test depression as the dependent variable, with pre-test depression, retrospective rumination change and EMA rumination change entered as predictors.

To reduce the influence of outliers in both the CARE and HMP samples, we winsorized any extreme values (Winsorize function in the DescTools R package). Data were analyzed using R version 4.2.2 (R Core Team, 2023).

Results

Reliability of self-report and EMA measures

CARE Samples: Since internal consistency is computed on the raw item scores and two of the adolescent samples (AMAS and ARCT) completed the CRSQ and the other (College) sample used the RSQ, we computed reliability separately for both measures. Reliability (internal consistency) for retrospective rumination was high at T1 (α = .89 for Adolescents and α = .93 for College) and T2 (α = .90 for Adolescents and α = .94 for College), but decreased for the residualized change score (ρ = .71 for Adolescents and .84 for College). Reliability (split-half) for EMA rumination was high (ρ = .89), but decreased substantially for the change score (ρ = .50; or using reliability estimate from MLM, ρ = .58). There were fewer EMA timepoints per subject in the CARE samples (Mean = 50.2, Median = 55.0, SD = 17.6) relative to the HMP sample (Mean = 89.9, Median = 98, SD = 23.1). Reliability estimates are often higher when there are more timepoints as the slopes are estimated with less error. We re-computed reliability for the change score excluding individuals with fewer than 40 timepoints, which increased the reliability coefficients from ρ = .50 to ρ = .61.

HMP Sample: Reliability (internal consistency) for retrospective rumination was high at T1 (α = .94) and T2 (α = .95). The residualized change score in this sample also retained high reliability (ρ = .90). Reliability (split-half) for EMA rumination was high (ρ = .96), but decreased for the change score (ρ = .77; =.87 using MLM-based reliability estimate). Note that the relatively high

Preprint JMIR Publications

residualized change reliabilities for the HMP and CARE College samples are due to the high reliabilities of the constituent T1 and T2 rumination scores (see above) and their relatively low intercorrelation (CARE College *r* = .66 and HMP *r* = .55 vs. CARE HMP = .69), implying more substantial variability in real change (see e.g., formula 7.10 of Crocker & Algina, 1986), p.149). See Supplement for additional analyses, including convergent and divergent validity tests for EMA, and the reliability of raw, rather than residualized, change scores which were similar.

Correlation between self-report and EMA measures

CARE Samples: Retrospective rumination at T1 was significantly positively correlated with mean EMA rumination over the subsequent 3 days (*r* = 0.35, *p* < .001; **Figure 1a**). Similarly, the correlation between retrospective rumination at T2 and EMA rumination over the prior 3 days was significant (*r* = 0.28, *p* < .001; **Figure 1b**). The magnitude of these correlations is conventionally considered to be in the medium range (Cohen, 1988).

Residualized T1-T2 change in retrospective rumination was not significantly correlated with change (slope) in EMA rumination (*r* = 0.03, *p* = .662; **Figure 1c**).

HMP Sample: Retrospective rumination at T1 was significantly positively correlated with mean EMA rumination over the first 3 days of EMA (*r* = .31, *p* = .003; **Figure 1d**). Similarly, the correlation between retrospective rumination at T2 and EMA rumination over the final 3 days of EMA was significant (*r* = .47, *p* < .001; **Figure 1e**). These correlations were medium-to-large in magnitude (Cohen, 1988).

Residualized T1-T2 change in retrospective repetitive negative thinking was not significantly correlated with change (slope) in EMA rumination (*r* = .20, *p* = .062; **Figure 1f**).

Comparing rumination improvement from self-report and EMA measures

CARE Samples: A significant group difference emerged for change in retrospective rumination, such that the mindfulness group exhibited significantly greater improvement relative to the mood monitoring control group (*d* = 0.37; *b* = -0.34, *SE* = 0.096, t(294) = -3.51, *p* < .001). In

Preprint JMIR Publications

contrast, there were no group differences in EMA rumination change (*d* = 0.14; *b* = -0.20, *SE* = 0.138, *p* = .140).

HMP Sample: There was a significant pre-post reduction in retrospective repetitive negative thinking in the HMP sample (*d* = 0.76, t(88) = 7.60, *p* < .001). EMA rumination decreased over time (*b* = -0.012, *SE* = 0.0050, *p* = .018). However, the associated effect size was small (*d* = 0.17).

Predictive validity of self-report and EMA measures

CARE Samples: Greater pre- to post-intervention improvement in both retrospective rumination (*b* = 0.04, *SE* = 0.017, *p* = .028) and EMA rumination (*b* = 0.03, *SE* = 0.014, *p* = .037) were significantly positively associated with improvement in depressive symptoms from baseline though follow-up.

HMP Sample: Similarly, greater pre- to post-intervention improvement in both retrospective rumination (*b* = 0.23, *SE* = 0.060, *p* < .001) and EMA rumination (*b* = 20.49, *SE* = 9.94, *p* = .042) were significantly positively associated with improvement in depression symptoms from baseline to post-test.

Discussion

Researchers are increasingly incorporating EMA into intervention studies in an effort to acquire a more fine-grained and ecologically valid assessment of change. EMA can provide high-resolution information about treatment-relevant processes (e.g., emotional, cognitive, and behavioral change), while minimizing memory biases relative to conventional (retrospective) self-report measures. However, repeated EMA surveys can be burdensome to participants, which could lead to poor compliance, careless responding (e.g., random or stereotyped responses to items), or a bias towards the enrollment of participants who are more motivated, thus limiting generalizability. In addition, the psychometric properties of EMA measures are often not well established prior to implementing them in a study (e.g., researchers often create new EMA items to assess a construct of interest, or adapt them from an existing self-report scale without evaluating or reporting their

Preprint JMIR Publications

psychometric properties) (Brose et al., 2020). In short, intervention researchers must carefully consider whether the benefits of implementing an EMA measure in a study outweigh potential costs and limitations. In the current study, we attempted to help address the question of whether EMA is worth including by combining data from four trials (total *N* = 412), all of which examined a meditation app and measured outcomes (change in rumination) via both conventional retrospective self-report questionnaires and EMA. We did so by evaluating four relevant criteria.

Criterion 1: Are EMA Measures of Rumination Reliable?

While the reliability of both retrospective rumination (at baseline and post-intervention) and mean EMA rumination scores were high, our findings revealed a decline in reliability when assessing change over time, especially for EMA. These results align with long-standing concerns about the unreliability of measures being compounded when measuring change (e.g., Cronbach & Furby, 1970; Willett, 1989). Reliability is generally defined as the ratio of systematic ("true") variance of interest to total observed variance. In addition to high measurement error, another reason for the low reliability in change scores is that true change variability may be low, either because of limited change, or because all participants change to a nearly equivalent degree (Williams & Zimmerman, 1996). Prior EMA studies have reported relatively lower reliability for change scores (e.g., Cranford et al., 2006; Dejonckheere et al., 2022; Haney et al., 2023). Critically, low reliability can result in biased inferences and a loss of statistical power, leading to an inaccurate estimation of the true effect in the population (Shrout & Lane, 2012). Consequently, when interpreting findings, it becomes extremely challenging to determine whether the results are trustworthy or whether they are attributable to measurement error, thereby reducing the likelihood of successfully replicating findings (Parsons et al., 2019). Despite the importance of establishing reliability, many EMA studies do not adequately report the psychometrics of their measures (Brose et al., 2020).

In theory, two approaches to improving the reliability (and likely the validity) of EMA measures are to increase the number of timepoints or add more items that capture the construct of

Preprint JMIR Publications

interest (Bolger & Laurenceau, 2013). In the current study, we relied on single item measures of EMA rumination, which may provide less coverage of the construct of interest relative to multi-item measures (Dejonckheere et al., 2022). However, it is important to note that increasing the number of items or assessment timepoints also increases the burden on participants and consequently may reduce compliance or increase careless responding (van Berkel et al., 2020). Therefore, researchers should balance the need to improve measurement (via more items or assessment timepoints) with a careful consideration of participant burden. In the current study, using single-item measures (a common practice in many EMA studies) also prevented us from calculating internal consistency, the method we employed to compute reliability for the conventional self-report measures. Consequently, we were unable to directly compare the reliabilities of the conventional self-report and EMA measures.

Similar to increasing the number of relevant items, averaging scores across multiple measurement timepoints is known to improve reliability (Bolger & Laurenceau, 2013; Brose et al., 2020). This approach aligns with our finding of high reliability for *mean* EMA rumination. For example, in the context of studying day-to-day changes in rumination, averaging data collected from multiple within-day occasions (as opposed to relying on a single measurement per day) to estimate rumination levels each day is expected to improve the precision of within-day rumination. In addition, the higher reliability for EMA change scores in the HMP relative to the CARE sample may be due to the larger number of timepoints collected per subject in the former sample improving the precision of slope estimates.

In summary, the low reliability of change scores (in particular for EMA measures of rumination) is concerning. It would be helpful to better understand the scope of this issue if researchers more consistently reported the reliability of change were it is evaluated, given that this is not common practice yet can have a profound impact on study findings (e.g., low reliability attenuating effect sizes). This can aid readers in evaluating the results of EMA findings in a given

Preprint JMIR Publications

study in light of potential psychometric limitations. The field may also benefit from additional innovative approaches to assessing the reliability of EMA measures (e.g., Dejonckheere et al., 2022).

Criterion 2: Correlation with Traditional Self-Report

Results revealed statistically significant, medium-to-large (*r*s = .28 to .47) correlations (Cohen, 1988) between rumination assessed via conventional, retrospective self-report and the mean of EMAs. However, there were no significant correlations between changes in retrospective rumination and change in EMA rumination (*r*s = .03 to .20). These results correspond with previous studies that have reported low-to-moderate correlations between trait and state measures of various constructs including anhedonia (Loas et al., 2009), affective lability (e.g., Anestis et al., 2010; Solhan et al., 2009) and personality traits (Rauthmann et al., 2019; Ringwald et al., 2022).

Several factors may account for the relatively modest association between conventional retrospective self-report and EMA measures of rumination (especially for the change scores). First, less than perfect reliability can attenuate associations between any two variables. Reliability estimates were particularly low for change scores which may explain (at least in part) the especially low correlation between those measures. Second, Kahneman's distinction between the "experiencing self" and the "remembering self" provides a useful framework for interpreting these results (Kahneman, 2011). According to this distinction, when participants are asked to recall their past experiences on conventional, retrospective self-report measures (in this study, asking about past ruminative thoughts), they may be especially influenced by the most intense (peak) and recent (end) levels of rumination (i.e., the so called “peak-end” bias), rather than equally weighting and simply averaging all timepoints, as reflected by the mean of repeated EMAs (Chajut et al., 2014; Kahneman et al., 1993; Redelmeier et al., 2003; Redelmeier & Kahneman, 1996). Third, whereas momentary EMA requires introspection of *current* states, retrospective self-report measures require participants to access autobiographical (episodic) memory and provide an average report of past experiences. The accuracy of these reports is limited by the extent to which past experiences have been accurately

Preprint JMIR Publications

encoded and consolidated in memory (Conner & Barrett, 2012; Miron-Shatz et al., 2009; Van den Bergh & Walentynowicz, 2016). Moreover, when recall timeframes are longer, individuals may shift from episodic memory to semantic memory, which includes more abstract beliefs or generalization about the self (Robinson & Clore, 2002; Van den Bergh & Walentynowicz, 2016). Thus, when asked to report on their typical tendency to ruminate on self-report measures, individuals' responses may be influenced by relevant self-concepts ("I tend to overthink", "I'm a worrier"), how they believe others view them, or how they would like to be perceived (Augustine & Larsen, 2012). In summary, in addition to reduced reliability, the relatively modest correlations between conventional retrospective self-report and EMA measures of rumination may be due to the fact that the two measurement approaches ultimately tap different conscious or functional "selves" (Conner & Barrett, 2012). As described below, the decision to include EMA and/or retrospective measures of change in an intervention study should be informed by which of these "selves" (experiencing vs. remembering self) the researchers want to assess and expect to change.

Criterion 3: Sensitivity to Change

Compared to conventional retrospective self-report, EMA measures detected smaller group (intervention vs. control) differences in rumination improvement (effect sizes *d* = 0.37 vs. *d* = 0.14, for self-report and EMA, respectively). EMA also showed less linear change in rumination over time in the intervention group (effect sizes *d* = 0.76 vs. *d* = 0.17, for self-report and EMA, respectively). These findings have important implications for the selection of outcome measures in intervention studies. They imply that a trial may yield very different findings (e.g., significant vs. non-significant differences in outcomes between the intervention and control group) based on whether a conventional, retrospective or EMA instrument was used as the outcome measure. This raises the critical question of which result is closer to the "ground truth." The larger effect size observed for retrospective self-report could be due to this measure being a more sensitive method for capturing changes in rumination. Although speculative, EMA may be less sensitive to detecting change due the

Preprint JMIR Publications

repeated nature of EMA surveys introducing unique forms of bias. For example, perhaps participants tend to align their responses with their (relatively recent) previous answers (a type of anchoring bias), potentially resulting in reduced detection of actual change. If a participant rated their rumination as a 4 on a 1-5 scale on the last EMA survey, an anchoring bias may lead them to report a similar score on the next survey, even if their "true" rumination score is now lower. Alternatively, EMA may provide a more *accurate* estimation of real change (which may in fact be modest), while the conventional retrospective self-report measure may overinflate estimates of change due to influences such as social desirability bias, regression to the mean or initial elevation bias (Shrout et al., 2018).

If, as discussed above, EMA and retrospective self-reports indeed capture related but somewhat different constructs, it is possible that the pattern of results we observed is due to the intervention differentially affecting each of these facets. For example, it may be that the intervention had a larger impact on one's episodic or semantic representations of being a ruminator (as measured by retrospective self-report) but had a more modest impact on actual moment-to-moment rumination (as measured by EMA). There may also be a causal relationship from the latter to the former. Namely, relatively modest improvements in day-to-day momentary rumination may have a proportionally large positive impact on one's self-concept of being a ruminator.

Even if changes in retrospective self-report are biased due to the limitations of memory, and there is little change in actual experience, these changes in retrospective recall may still be meaningful. Research has shown that future behaviors (e.g., whether to return for a colonoscopy screening or end a romantic relationship) are *better* predicted by retrospective measures of experience than actual momentary experiences (Shiffman et al., 2008). In addition, improvement in summary, retrospective self-reports (e.g., related to one's self-concept and tendency to engage in repetitive negative thought) may, of course, be very personally meaningful to the individual who underwent the intervention. Therefore, if the goal is to understand what individuals experienced in

Preprint JMIR Publications

the moment, EMA arguably provides a more accurate measure. However, if the focus is on individuals' global impressions of their experiences and themselves, then retrospective self-report measures may be more valuable.

Criterion 4: Predicting Change in Depressive Symptoms

Finally, results of the current study revealed that both retrospective and EMA measures of rumination change predicted improvement in depressive symptoms, with each measure contributing significantly beyond the effects of the other (i.e., evidence of the incremental predictive validity of each measure). These findings suggest that these two methods assess meaningfully distinct aspects of rumination, both of which hold clinically important information. Therefore, if feasible, combining retrospective and EMA data could provide a more comprehensive picture of clinically meaningful improvement over time. It is also important to highlight that there are of course other reasons why a researcher may want to include EMA measures in an intervention study (e.g., to test whether emotion or symptom dynamics change over time)(Piccirillo & Rodebaugh, 2019).

Strengths and Limitations

There are several notable strengths to this study. First, we combined data across 4 trials to increase sample size (total *N* = 412). Second, each of the studies examined a meditation app and measured outcomes via both retrospective self-report and EMA. Finally, we evaluated several metrics of reliability, validity, sensitivity to change and incremental predictive validity to examine the value of EMA measures in an intervention study. At the same time, the study had several limitations. First, with regards to criterion 4, a well-validated clinical interview of depressive symptoms would have been preferable to a self-report instrument to serve as the "ground truth" of change during the intervention. Second, in contrast to the CARE samples which had follow-up timepoints, the HMP sample analyses for criterion 4 (*Predicting Change in Depressive Symptoms*) were limited in testing the relation between change in rumination and depressive symptoms over the same timeframe (i.e., pre- to post-treatment), which is problematic (Shahar, 2009) relative to lagged analyses testing whether early change in rumination prospectively predicts subsequent depressive symptom improvement. Third, we only focused on rumination. The extent to which our findings generalize to other common outcome measures (e.g., comparing conventional self-report and EMA measures of change in depression or anxiety symptoms) is unknown. Fourth, the fact that the items in the retrospective and EMA measures of rumination were not identical likely attenuated correlations. Fifth, it would have been preferable to have a multi-item rather than single-item measure of EMA rumination which would have allowed us to compute internal consistency and multilevel reliability estimates (Dejonckheere et al., 2022). On the other hand, given the burden of EMA, single-item measures are *very* common and thus our study may generalize to the existing literature. Sixth, it may be that mindfulness training (or perhaps even repeatedly answering rumination questions in the control group) could have shifted how participants interpret and respond to the rumination prompts (Fried et al., 2016; McNeish et al., 2021; Van Dam et al., 2009). For example, on the topic of the influence of meditation on item interpretation and response patterns, there is evidence of differential item functioning among meditators and non-meditators on mindfulness measures (Van Dam et al., 2009). Thus, it may be that meditation practice (e.g., cultivating non-judgmental awareness of internal experience) shifts the interpretation of mindfulness items, which could also apply to the rating of negative mental states assessed by rumination items (e.g., see above for our concern about the focusing on emotions item). However, the use of only one item for measurement in EMA, combined with the relatively low sample sizes precluded our ability to meaningfully evaluate longitudinal measurement invariance. Finally, the fact that the included samples/studies differed in patient characteristics (e.g., age) and methods (e.g., different measures of rumination) could be considered a limitation. On the other hand, it allowed us to test the consistency of findings (e.g., reliability of and correlations between retrospective and EMA measures) across these sample/study differences (which, overall, were quite consistent which provides some evidence, albeit limited, for the generalizability of findings).

Preprint
JMIR Publications

Despite these limitations, the results of the present study provide insight into the utility of conventional self-report and EMA measures of rumination that are relevant considerations for researchers when designing studies and interpretating results. First, the reliability of change in rumination over the course of the intervention were relatively low, especially for EMA. This is concerning given that change over time is precisely what intervention researchers are most interested in. Increasing the number of EMA assessment timepoints and using multi-item scales, while being mindful of not overburdening participants, may improve the precision with which change over time is estimated. The frequently observed low reliability of change is often presented as a concern with the measurement of change more generally (Bandalos, 2018; pp. 201-205). However, beyond the reliability of measures at each timepoint, the most critical factor is the correlation between measures across timepoints, which can be viewed as an index of the variability in change. As this will be a feature of both study and population conditions that will not be consistent across studies and not under the direct control of the investigator, we concur with prior commentaries on the issue that reliability of change scores should be evaluated on a case-by-case basis rather than dismissed out-of-hand (Collins, 1996; Thomas & Zumbo, 2012; Trafimow, 2015).

Second, conventional self-report and EMA measures of rumination were modestly correlated at a given timepoint with no significant correlation between changes in retrospective and EMA measures of rumination over time. Despite this, changes in both measures of rumination predicted decreased depressive symptoms at follow up. This suggests that conventional self-report and EMA measures of rumination are not redundant. Rather, each measure is capturing distinct and clinically meaningful facets of rumination (i.e., remembered vs. experienced rumination). Using both conventional self-report and EMA measures of rumination in combination, when feasible, may provide a more comprehensive picture of clinical change. It can be appreciated that the tendency for conventional self-report to administer many items at few time points, and the tendency for EMA to administer fewer items are many time points, also provide complementary designs that minimize the

Preprint JMIR Publications

effects of different sources of error (in the case of self-report, item-related error, and in the case of EMA, time-related error), which may be useful to quantify using generalizability theory techniques, and that can be incorporated into a multilevel reliability analysis. There is value for future research to carefully consider such complexities related to the psychometric properties of EMA measures of rumination and ideally to ultimately standardize measures of rumination in the EMA literature.

Preprint JMIR Publications

Author Contributions

Christian A. Webb: Conceptualization; Formal analysis; Visualization; Writing – original draft; Writing – review & editing

Lori M. Hilt: Conceptualization; Funding Acquisition; Investigation; Data curation; Writing – original draft; Writing – review & editing

Caroline M. Swords: Project Administration: Investigation; Data curation; Writing – original draft; Writing – review & editing

Daniel M. Bolt: Conceptualization; Formal analysis; Writing – original draft; Writing – review & editing

Hadar Fisher: Conceptualization; Writing – original draft; Writing – review & editing

Simon B. Goldberg: Conceptualization; Funding Acquisition; Formal analysis; Writing – original draft; Writing – review & editing

Acknowledgments

This research was supported by NCCIH Grant K23AT010879 (SBG), the American Psychological Foundation (LMH) and NIMH Grant R15MH116303 (LMH). The first author (CAW) was partially supported by NIMH R01MH116969, NCCIH R01AT011002, the Tommy Fuss Fund and a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funders.

Conflicts of Interest

Dr. Webb has received consulting fees from King & Spalding law firm for work unrelated to this publication.

References:

Aalto, M., Alho, H., Halme, J. T., & Seppä, K. (2009). AUDIT and its abbreviated versions in

Preprint JMIR Publications

detecting heavy and binge drinking in a general population survey. *Drug and Alcohol Dependence*, *103*(1), 25–29. https://doi.org/10.1016/j.drugalcdep.2009.02.013

Abela, J. R. Z., Brozina, K., & Haigh, E. P. (2002). An examination of the response styles theory of depression in third- and seventh-grade children: A short-term longitudinal study. *Journal of Abnormal Child Psychology*, *30*(5), 515–527. https://doi.org/10.1023/A:1019873015594

Andresen, E. M., Malmgren, J. A., Carter, W. B., & Patrick, D. L. (1994). Screening for depression in well older adults: Evaluation of a short form of the CES-D (Center for Epidemiologic Studies Depression Scale). *American Journal of Preventive Medicine*, *10*(2), 77–84.

Anestis, M. D., Selby, E. A., Crosby, R. D., Wonderlich, S. A., Engel, S. G., & Joiner, T. E. (2010). A comparison of retrospective self-report versus ecological momentary assessment measures of affective lability in the examination of its relationship with bulimic symptomatology. *Behaviour Research and Therapy*, *48*(7), 607–613.

Augustine, A. A., & Larsen, R. J. (2012). Is a trait really the mean of states? *Journal of Individual Differences*.

Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences* (1st edition). The Guilford Press.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Manual for the Beck depression inventory-II. *San Antonio, TX: Psychological Corporation*, 1–82.

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research* (1 edition). The Guilford Press.

Broderick, J. E., Schwartz, J. E., Vikingstad, G., Pribbernow, M., Grossman, S., & Stone, A. A. (2008). The accuracy of pain and fatigue items across different reporting periods. *PAIN*, *139*(1), 146. https://doi.org/10.1016/j.pain.2008.03.024

Brose, A., Schmiedek, F., Gerstorf, D., & Voelkle, M. C. (2020). The measurement of within-person affect variation. *Emotion*, *20*(4), 677–699.

Preprint JMIR Publications

Cerino, E. S., Schneider, S., Stone, A. A., Sliwinski, M. J., Mogle, J., & Smyth, J. M. (2022). Little evidence for consistent initial elevation bias in self-reported momentary affect: A coordinated analysis of ecological momentary assessment studies. *Psychological Assessment*, *34*(5), 467–482. https://doi.org/10.1037/pas0001108

Chajut, E., Caspi, A., Chen, R., Hod, M., & Ariely, D. (2014). In Pain Thou Shalt Bring Forth Children: The Peak-and-End Rule in Recall of Labor Pain. *Psychological Science*, *25*(12), 2266–2271. https://doi.org/10.1177/0956797614551004

Chen, E., Zeltzer, L. K., Craske, M. G., & Katz, E. R. (2000). Children's memories for painful cancer treatment procedures: Implications for distress. *Child Development*, *71*(4), 933–947. https://doi.org/10.1111/1467-8624.00200

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 edition). Routledge.

Collins, L. M. (1996). Is Reliability Obsolete? A Commentary on "Are Simple Gain Scores Obsolete?" *Applied Psychological Measurement*, *20*(3), 289–292. https://doi.org/10.1177/014662169602000308

Conner, T. S., & Barrett, L. F. (2012). Trends in ambulatory self-report: The role of momentary experience in psychosomatic medicine. *Psychosomatic Medicine*, *74*(4), 327–337. https://doi.org/10.1097/PSY.0b013e3182546f18

Craighead, W. E., Curry, J. F., & Ilardi, S. S. (1995). Relationship of Children's Depression Inventory factors to major depression among adolescents. *Psychological Assessment*, *7*(2), 171–176. https://doi.org/10.1037/1040-3590.7.2.171

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

Preprint JMIR Publications

change reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917–929.

Crocker, L. M., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Holt, Rinehart, and Winston.

Cronbach, L. J., & Furby, L. (1970). How we should measure" change": Or should we? *Psychological Bulletin*, *74*(1), 68–80.

Dahl, C. J., Wilson-Mendenhall, C. D., & Davidson, R. J. (2020). The plasticity of well-being: A training-based framework for the cultivation of human flourishing. *Proceedings of the National Academy of Sciences*, *117*(51), 32197–32206.

Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment*, *34*(12), 1138–1154.

Ehring, T., Zetsche, U., Weidacker, K., Wahl, K., Schönfeld, S., & Ehlers, A. (2011). The Perseverative Thinking Questionnaire (PTQ): Validation of a content-independent measure of repetitive negative thinking. *Journal of Behavior Therapy and Experimental Psychiatry*, *42*(2), 225–232.

Faul, L., & LaBar, K. S. (2022). Mood-congruent memory revisited. *Psychological Review*. https://doi.org/10.1037/rev0000394

Feingold, A. (2009). Effect Sizes for Growth-Modeling Analysis for Controlled Clinical Trials in the Same Metric as for Classical Analysis. *Psychological Methods*, *14*(1), 43–53. https://doi.org/10.1037/a0014699

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time . . . Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, *28*(11), 1354–1367. https://doi.org/10.1037/pas0000275

Friedberg, F., & Sohl, S. J. (2008). Memory for fatigue in chronic fatigue syndrome: The relation

Preprint JMIR Publications

between weekly recall and momentary ratings. *International Journal of Behavioral Medicine*, *15*(1), 29–33. https://doi.org/10.1007/BF03003071

Geng, X., Chen, Z., Lam, W., & Zheng, Q. (2013). Hedonic Evaluation over Short and Long Retention Intervals: The Mechanism of the Peak–End Rule. *Journal of Behavioral Decision Making*, *26*(3), 225–236. https://doi.org/10.1002/bdm.1755

Giske, L., Sandvik, L., & Røe, C. (2010). Comparison of daily and weekly retrospectively reported pain intensity in patients with localized and generalized musculoskeletal pain. *European Journal of Pain (London, England)*, *14*(9), 959–965. https://doi.org/10.1016/j.ejpain.2010.02.011

Hamilton, M. (1960). A Rating Scale for Depression. *Journal of Neurology, Neurosurgery & Psychiatry*, *23*(1), 56–62. https://doi.org/10.1136/jnnp.23.1.56

Haney, A. M., Fleming, M. N., Wycoff, A. M., Griffin, S. A., & Trull, T. J. (2023). Measuring affect in daily life: A multilevel psychometric evaluation of the PANAS-X across four ecological momentary assessment samples. *Psychological Assessment*, *35*(6), 469–483.

Hernandez, R., Schneider, S., Pinkham, A. E., Depp, C. A., Ackerman, R., Pyatak, E. A., Badal, V. D., Moore, R. C., Harvey, P. D., Funsch, K., & Stone, A. A. (2024). Comparisons of Self-Report With Objective Measurements Suggest Faster Responding but Little Change in Response Quality Over Time in Ecological Momentary Assessment Studies. *Assessment*, 10731911241245793. https://doi.org/10.1177/10731911241245793

Hilt, L. M., & Pollak, S. D. (2012). Getting out of rumination: Comparison of three brief interventions in a sample of youth. *Journal of Abnormal Child Psychology*, *40*(7), 1157–1165. https://doi.org/10.1007/s10802-012-9638-3

Hilt, L. M., Sladek, M. R., Doane, L. D., & Stroud, C. B. (2017). Daily and trait rumination: Diurnal cortisol patterns in adolescent girls. *Cognition and Emotion*, *31*(8), 1757–1767.

Hilt, L. M., & Swords, C. M. (2021). Acceptability and Preliminary Effects of a Mindfulness Mobile

Preprint JMIR Publications

Application for Ruminative Adolescents. *Behavior Therapy*, *6*(52), 1339–1350. https://doi.org/10.1016/j.beth.2021.03.004

Hilt, L. M., Swords, C. M., & Webb, C. A. (2023). Randomized Controlled Trial of a Mindfulness Mobile Application for Ruminative Adolescents. *Journal of Clinical Child & Adolescent Psychology*, *0*(0), 1–14. https://doi.org/10.1080/15374416.2022.2158840

Jaso, B. A., Kraus, N. I., & Heller, A. S. (2022). Identification of careless responding in ecological momentary assessment research: From posthoc analyses to real-time data monitoring. *Psychological Methods*, *27*(6), 958–981. https://doi.org/10.1037/met0000312

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Strauss, Giroux.

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

Klein, D. N., Dougherty, L. R., & Olino, T. M. (2005). Toward Guidelines for Evidence-Based Assessment of Depression in Children and Adolescents. *Journal of Clinical Child & Adolescent Psychology*, *34*(3), 412–432. https://doi.org/10.1207/s15374424jccp3403_3

Kovacs, M. (1992). *Children's Depression Inventory Manual*. Multi-Health Systems.

Liu, S., Kuppens, P., & Bringmann, L. (2021). On the Use of Empirical Bayes Estimates as Measures of Individual Traits. *Assessment*, *28*(3), 845–857. https://doi.org/10.1177/1073191119885019

Loas, G., Monestes, J. L., Ingelaere, A., Noisette, C., & Herbener, E. S. (2009). Stability and relationships between trait or state anhedonia and schizophrenic symptoms in schizophrenia: A 13-year follow-up study. *Psychiatry Research*, *166*(2–3), 132–140.

McNeish, D., Mackinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in Intensive Longitudinal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 807–822. https://doi.org/10.1080/10705511.2021.1915788

Miron-Shatz, T., Stone, A., & Kahneman, D. (2009). Memories of yesterday's emotions: Does the

Preprint JMIR Publications

valence of experience affect the memory-experience gap? *Emotion (Washington, D.C.)*, *9*(6), 885–891. https://doi.org/10.1037/a0017823

Moberly, N. J., & Watkins, E. R. (2008). Ruminative self-focus and negative affect: An experience sampling study. *Journal of Abnormal Psychology*, *117*(2), 314.

Mofsen, A. M., Rodebaugh, T. L., Nicol, G. E., Depp, C. A., Miller, J. P., & Lenze, E. J. (2019). When All Else Fails, Listen to the Patient: A Viewpoint on the Use of Ecological Momentary Assessment in Clinical Trials. *JMIR Mental Health*, *6*(5), e11845. https://doi.org/10.2196/11845

Moore, R. C., Depp, C. A., Wetherell, J. L., & Lenze, E. J. (2016). Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults. *Journal of Psychiatric Research*, *75*, 116–123. https://doi.org/10.1016/j.jpsychires.2016.01.011

National Center on Intensive Intervention. (2014). *Homepage | NCII*. https://intensiveintervention.org/sites/default/files/APM_FAQs_2014.pdf

Neubauer, A. B., Scott, S. B., Sliwinski, M. J., & Smyth, J. M. (2020). How was your day? Convergence of aggregated momentary and retrospective end-of-day affect ratings across the adult life span. *Journal of Personality and Social Psychology*, *119*(1), 185–203. https://doi.org/10.1037/pspp0000248

Noel, M., Rabbitts, J. A., Tai, G. G., & Palermo, T. M. (2015). Remembering pain after surgery: A longitudinal examination of the role of pain catastrophizing in children’s and parents’ recall. *Pain*, *156*(5), 800–808. https://doi.org/10.1097/j.pain.0000000000000102

Nolen-Hoeksema, S. (1991). Responses to depression and their effects on the duration of depressive episodes. *Journal of Abnormal Psychology*, *100*(4), 569.

Nolen-Hoeksema, S., & Morrow, J. (1991). A prospective study of depression and posttraumatic stress symptoms after a natural disaster: The 1989 Loma Prieta Earthquake. *Journal of*

Preprint JMIR Publications

Personality and Social Psychology, *61*(1), 115–121.

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395.

Peterson, C. B., Engel, S. G., Crosby, R. D., Strauman, T., Smith, T. L., Klein, M., Crow, S. J., Mitchell, J. E., Erickson, A., Cao, L., Bjorlie, K., & Wonderlich, S. A. (2020). Comparing integrative cognitive-affective therapy and guided self-help cognitive-behavioral therapy to treat binge-eating disorder using standard and naturalistic momentary outcome measures: A randomized controlled trial. *International Journal of Eating Disorders*, *53*(9), 1418–1427. https://doi.org/10.1002/eat.23324

Piccirillo, M. L., & Rodebaugh, T. L. (2019). Foundations of idiographic methods in psychology and applications for psychotherapy. *Clinical Psychology Review*. https://doi.org/10.1016/j.cpr.2019.01.002

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & Group, P. C. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment*, *18*(3), 263–283.

Poznanski, E. O., & Mokros, H. B. (1996). *Children's depression rating scale, revised (CDRS-R)*. Western Psychological Services Los Angeles.

Rauthmann, J. F., Horstmann, K. T., & Sherman, R. A. (2019). Do self-reported traits and aggregated states capture the same thing? A nomological perspective on trait-state homomorphy. *Social Psychological and Personality Science*, *10*(5), 596–611.

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, *66*(1), 3–8. https://doi.org/10.1016/0304-3959(96)02994-6

Preprint JMIR Publications

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

Ringwald, W. R., Manuck, S. B., Marsland, A. L., & Wright, A. G. (2022). Psychometric evaluation of a Big Five personality state scale for intensive longitudinal studies. *Assessment*, *29*(6), 1301–1319.

Rinner, M. T. B., Meyer, A. H., Mikoteit, T., Hoyer, J., Imboden, C., Hatzinger, M., Bader, K., Lieb, R., Miché, M., Wersebe, H., & Gloster, A. T. (2019). General or specific? The memory–experience gap for individuals diagnosed with a major depressive disorder or a social phobia diagnosis, and individuals without such diagnoses. *Memory*, *27*(9), 1194–1203. https://doi.org/10.1080/09658211.2019.1640252

Robinson, M. D., & Clore, G. L. (2002). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology*, *83*(1), 198–215.

Roelofs, J., Muris, P., Huibers, M., Peeters, F., & Arntz, A. (2006). On the measurement of rumination: A psychometric evaluation of the ruminative response scale and the rumination on sadness scale in undergraduates. *Journal of Behavior Therapy and Experimental Psychiatry*, *37*(4), 299–313. https://doi.org/10.1016/j.jbtep.2006.03.002

Ruscio, A. M., Gentes, E. L., Jones, J. D., Hallion, L. S., Coleman, E. S., & Swendsen, J. (2015). Rumination predicts heightened responding to stressful life events in major depressive disorder and generalized anxiety disorder. *Journal of Abnormal Psychology*, *124*(1), 17–26. https://doi.org/10.1037/abn0000025

Schalet, B. D., Cook, K. F., Choi, S. W., & Cella, D. (2014). Establishing a Common Metric for Self-Reported Anxiety: Linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. *Journal of Anxiety Disorders*, *28*(1), 88–96. https://doi.org/10.1016/j.janxdis.2013.11.006

Schneider, S., Hernandez, R., Junghaenel, D. U., Orriens, B., Lee, P.-J., & Stone, A. A. (2024).

Response times in Ecological Momentary Assessment (EMA): Shedding light on the response process with a drift diffusion model. *Current Psychology*, *43*(7), 5868–5886. https://doi.org/10.1007/s12144-023-04773-0

Shahar, E. (2009). Evaluating the effect of change on change: A different viewpoint. *Journal of Evaluation in Clinical Practice*, *15*(1), 204–207. https://doi.org/10.1111/j.1365-2753.2008.00983.x

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, *4*, 1–32.

Shrout, P. E., & Lane, S. P. (2012). Psychometrics. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 302–320). The Guilford Press.

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences*, *115*(1), E15–E23. https://doi.org/10.1073/pnas.1712277115

Solhan, M. B., Trull, T. J., Jahng, S., & Wood, P. K. (2009). Clinical assessment of affective instability: Comparing EMA indices, questionnaire reports, and retrospective recall. *Psychological Assessment*, *21*(3), 425.

Stone, A. A., Broderick, J. E., Shiffman, S. S., & Schwartz, J. E. (2004). Understanding recall of weekly pain from a momentary assessment perspective: Absolute agreement, between- and within-person consistency, and judged change in weekly pain. *Pain*, *107*(1–2), 61–69. https://doi.org/10.1016/j.pain.2003.09.020

Stone, A. A., & Shiffman, S. S. (2010). Ecological Validity for Patient Reported Outcomes. In A. Steptoe (Ed.), *Handbook of Behavioral Medicine: Methods and Applications* (pp. 99–112). Springer. https://doi.org/10.1007/978-0-387-09488-5_8

Storch, E. A., Roberti, J. W., & Roth, D. A. (2004). Factor structure, concurrent validity, and internal

Preprint JMIR Publications

consistency of the Beck Depression Inventory-Second Edition in a sample of college students. *Depression and Anxiety*, *19*(3), 187–189. https://doi.org/10.1002/da.20002

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

Thomas, D. R., & Zumbo, B. D. (2012). Difference Scores From the Point of View of Reliability and Repeated-Measures ANOVA: In Defense of Difference Scores for Data Analysis. *Educational and Psychological Measurement*, *72*(1), 37–43. https://doi.org/10.1177/0013164411409929

Trafimow, D. (2015). A defense against the alleged unreliability of difference scores. *Cogent Mathematics*, *2*(1), 1064626. https://doi.org/10.1080/23311835.2015.1064626

Tulving, E. (2002). Episodic Memory: From Mind to Brain. *Annual Review of Psychology*, *53*(1), 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

van Berkel, N., Goncalves, J., Hosio, S., Sarsenbayeva, Z., Velloso, E., & Kostakos, V. (2020). Overcoming compliance bias in self-report studies: A cross-study analysis. *International Journal of Human-Computer Studies*, *134*, 1–12.

Van Dam, N. T., Earleywine, M., & Danoff-Burg, S. (2009). Differential item function across meditators and non-meditators on the Five Facet Mindfulness Questionnaire. *Personality and Individual Differences*, *47*(5), 516–521. https://doi.org/10.1016/j.paid.2009.05.005

Van den Bergh, O., & Walentynowicz, M. (2016). Accuracy and bias in retrospective symptom reporting. *Current Opinion in Psychiatry*, *29*(5), 302–308. https://doi.org/10.1097/YCO.0000000000000267

Van Norman, E. R., & Parker, D. C. (2018). A Comparison of Split-Half and Multilevel Methods to Assess the Reliability of Progress Monitoring Outcomes. *Journal of Psychoeducational Assessment*, *36*(6), 616–627. https://doi.org/10.1177/0734282917696936

Preprint JMIR Publications

Walentynowicz, M., Schneider, S., & Stone, A. A. (2018). The effects of time frames on self-report. *PLOS ONE*, *13*(8), e0201655. https://doi.org/10.1371/journal.pone.0201655

Webb, C. A., Israel, E. S., Belleau, E., Appleman, L., Forbes, E. E., & Pizzagalli, D. A. (2021). Mind-Wandering in Adolescents Predicts Worse Affect and Is Linked to Aberrant Default Mode Network–Salience Network Connectivity. *Journal of the American Academy of Child & Adolescent Psychiatry*, *60*(3), 377–387. https://doi.org/10.1016/j.jaac.2020.03.010

Webb, C. A., Murray, L., Tierney, A. O., & Gates, K. M. (2023). Dynamic processes in behavioral activation therapy for anhedonic adolescents: Modeling common and patient-specific relations. *Journal of Consulting and Clinical Psychology*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/ccp0000830

Webb, C. A., Swords, C. M., Lawrence, H. R., & Hilt, L. M. (2022). Which adolescents are well-suited to app-based mindfulness training? A randomized clinical trial and data-driven approach for personalized recommendations. *Journal of Consulting and Clinical Psychology*, *90*, 655–669. https://doi.org/10.1037/ccp0000763

Webb, C. A., Swords, C. M., Murray, L. M., & Hilt, L. M. (2021). App-based Mindfulness Training for Adolescent Rumination: Predictors of Immediate and Cumulative Benefit. *Mindfulness*, *12*, 2498–2509. https://doi.org/10.1007/s12671-021-01719-0

Wichers, M. C., Barge-Schaapveld, D. Q. C. M., Nicolson, N. A., Peeters, F., de Vries, M., Mengelers, R., & van Os, J. (2009). Reduced Stress-Sensitivity or Increased Reward Experience: The Psychological Mechanism of Response to Antidepressant Medication. *Neuropsychopharmacology*, *34*(4), Article 4. https://doi.org/10.1038/npp.2008.66

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement*, *49*(3), 587–602.

Williams, R. H., & Zimmerman, D. W. (1982). The Comparative Reliability of Simple and

Preprint JMIR Publications

Residualized Difference Scores. *The Journal of Experimental Education*, *51*(2), 94–97. https://doi.org/10.1080/00220973.1982.11011846

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*(1), 59–69.

Zurbriggen, C. L. A., Jendryczko, D., & Nussbeck, F. W. (2021). Rosy or blue? Change in recall bias of students' affective experiences during early adolescence. *Emotion (Washington, D.C.)*, *21*(8), 1637–1649. https://doi.org/10.1037/emo0001031

Preprint JMIR Publications

Table 1

Demographic Characteristics of Youth Participants at Baseline

Characteristics	Sample 1 (*n* = 80)		Sample 2 (*n* = 152)		Sample 3 (*n* = 88)		Sample 4 (*n* = 92)	
	n	%	*n*	%	*n*	%	*n*	%
Gender								
Male	43	53.8%	63	41.4%	26	29.5%	13	14.1%
Female	36	45.0%	89	58.6%	58	65.9%	75	81.5%
Non-binary	0	0.0%	0	0.0%	4	4.5%	4	4.3%
Chose not to answer	1	1.3%	0	0.0%	0	0.0%	0	0.0%
Race								
White	69	86.3%	125	82.2%	59	67.0%	64	69.6%
Black/African American	1	1.3%	5	3.3%	8	9.1%	3	3.3%
Asian	0	0.0%	3	2.0%	18	20.5%	18	19.6%
Native American/Alaskan Native	2	2.5%	0	0.0%	0	0.0%	2	2.2%
Native Hawaiian/Pacific Islander	0	0.0%	1	0.7%	0	0.0%	0	0.0%
Multiracial	1	1.3%	16	10.5%	0	0.0%	4	4.3%
Chose not to answer	7	8.8%	2	1.3%	3	3.4%	1	1.1%
Ethnicity								
Hispanic	3	3.8%	16	10.5%	10	11.4%	8	8.7%
Non-Hispanic	75	93.8%	136	89.5%	78	88.6%	83	90.2%
Chose not to answer	2	2.5%	0	0.0%	0	0.00%	1	1.1%

Preprint JMIR Publications

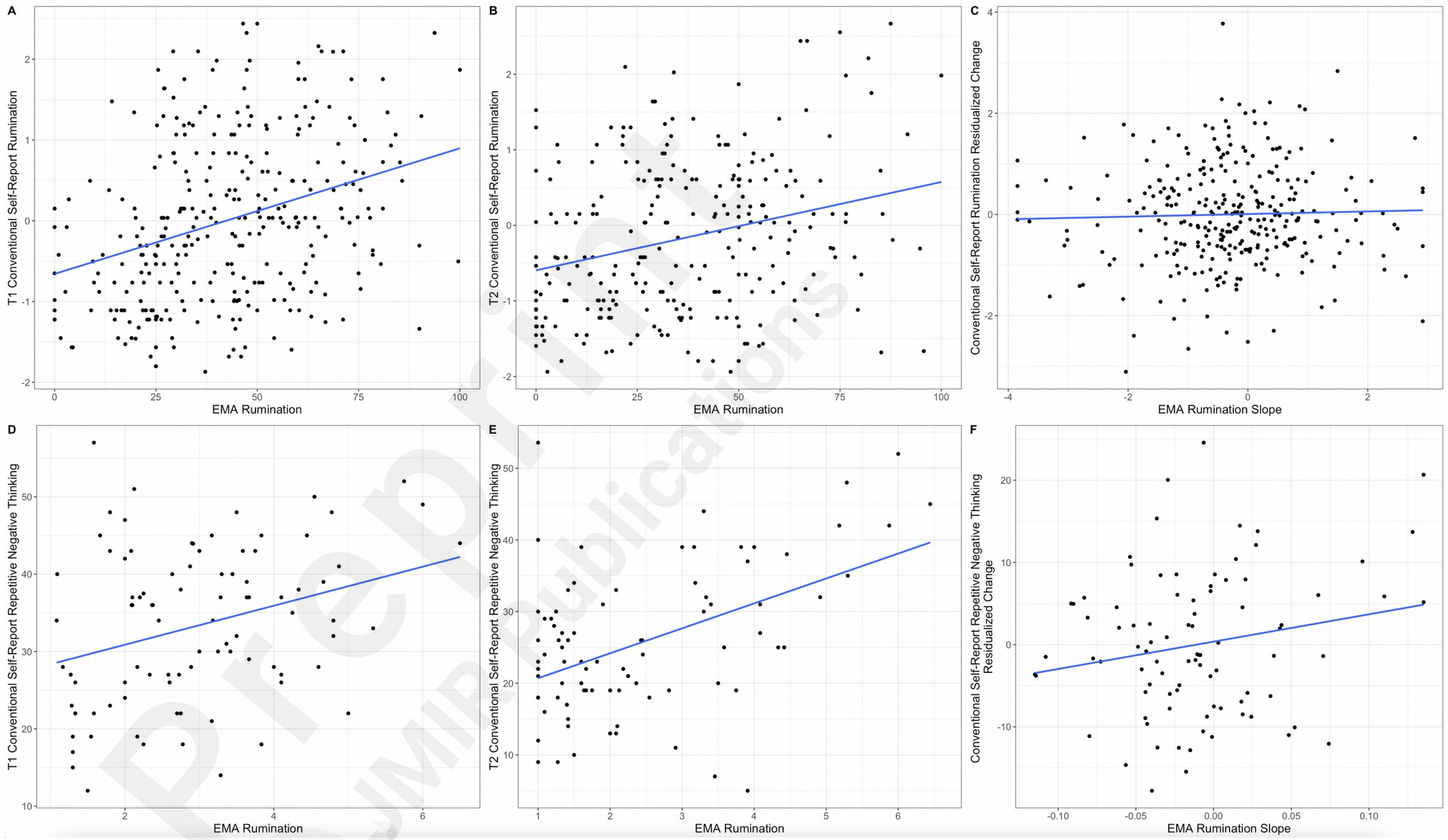
Figure Captions

Figure 1. Scatterplot of the association between: (1) conventional retrospective rumination at T1 and EMA rumination over the subsequent 3 days for the CARE (**Panel A**) and HMP (**Panel D**) samples; (2) conventional retrospective rumination at T2 and EMA rumination over the prior 3 days for the CARE (**Panel B**) and HMP (**Panel E**) samples; (3) residualized T1-T2 change in retrospective rumination and change (slope) in EMA rumination for the CARE (**Panel C**) and HMP (**Panel F**) samples.

Preprint JMIR Publications

ECOLOGICAL MOMENTARY ASSESSMENT MEASURES

Figure 1.



Preprint JMIR Publications

Data Transparency Statement: Data were drawn from four clinical trials. Primary outcome data has been published for 3 of these trials. Primary results from the fourth trial have not yet been published. The current study aggregates data across these four trials to evaluates the psychometric properties of EMA vs. conventional self-report measures or improvement in rumination. This question has not yet been addressed within these studies.

Preprint JMIR Publications

Preprint JMIR Publications

Supplementary Files

Multimedia Appendixes

Supplementary results.
URL: http://asset.jmir.pub/assets/3fc47d37fa28f629b60691ef3023f18a.docx

Preprint
JMIR Publications