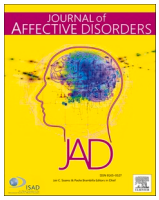
Journal of Affective Disorders 391 (2025) 119788

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad



Research paper

Evaluating psychometric validity of ecological momentary assessment for youth irritability

Eleanor G. Hansen [a], Reut Naim [b,c], Lauren M. Henry [a], Katharina Kircanski [a], Daniel S. Pine [a], Melissa A. Brotman [a,*], Meghan E. Byrne [a]

[a] Emotion and Development Branch, National Institute of Mental Health Intramural Research Program, Bethesda, MD, USA
[b] School of Psychological Sciences, Tel-Aviv University, Tel-Aviv, Israel
[c] Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv, Israel

ARTICLE INFO

Keywords:
Ecological momentary assessment
Irritability
Youth
Psychometrics
Validity

ABSTRACT

Background: Irritability is a transdiagnostic symptom in youth, leading to long-term adverse consequences. Ecological momentary assessment (EMA), or naturalistic clinical phenotyping, can quantify real-world experiences and affective dynamics of irritability in vivo and may be less contaminated by biases that impact retrospective report measures. However, to date, no research examines the psychometric properties of EMA measures of irritability.

Methods: The current study assesses EMA data from $N = 49$ youth receiving treatment for clinically impairing irritability (Mage = 10.63, 36.73 % female). Analyses evaluate two irritability EMA items and one positive-affect EMA item for within and between person variability, internal consistency, test-retest reliability, and convergent and discriminant validity.

Results: EMA items demonstrate acceptable psychometric properties including acceptable variability, consistency, reliability, and convergent and discriminant validity. Comparisons to previous work in anxious and healthy youth are discussed.

Limitations: Study limitations include participants' concurrent involvement in treatment and exclusion of outburst-related EMA measures from study analyses.

Conclusion: These results may facilitate future research with irritability EMA items in clinical populations; future work should validate EMA items psychometrically before use in clinical trials.

1. Introduction

Irritability, or an increased proneness to anger relative to peers (Brotman et al., 2017), is a common reason that youth and families seek psychiatric care (Collishaw et al., 2010; Evans et al., 2023a, 2023b) and predicts long-term adverse clinical, academic, and socioeconomic outcomes (Stringaris et al., 2009). Irritability is a transdiagnostic clinical phenomenon, implicated in multiple psychiatric diagnoses (Leibenluft, 2017). Irritability involves elevated propensity to frustration and displays of anger following blocked goal attainment (Leibenluft, 2017). Ecological momentary assessment (EMA) (Shiffman et al., 2008) leverages smartphone technology to repeatedly assess experiences in the moments they occur; it can quantify youths' real-world experiences of irritability (Naim et al., 2021) in ways that have advantages over retrospective report (Trull and Ebner-Priemer, 2013). However, EMA items for irritability have yet to be evaluated psychometrically to investigate internal validity and reliability for use within clinical research and trials (Byrne et al., 2023; Forkmann et al., 2018; Mestdagh and Dejonckheere, 2021).

EMA has grown in popularity in youth psychopathology research in the last decade (Henry et al., 2024) and offers many advantages compared to traditional self-report, parent-report, and clinician-report measures to assess youth irritability. Assessments occur within the participant's natural environment, enhancing ecological validity (Verhagen et al., 2016) and decreasing the probability of retrospective recall bias (Trull and Ebner-Priemer, 2013). Hypothesized components of irritability include temporal and contextual components (e.g., pervasive mood states, triggered outbursts) that may be better characterized by EMA methods than retrospective report (Evans et al., 2023a, 2023b; Naim et al., 2021; Naim et al., 2022). Further, affective dynamics

* Corresponding author.
E-mail address: brotmanm@mail.nih.gov (M.A. Brotman).

https://doi.org/10.1016/j.jad.2025.119788
Received 7 January 2025; Received in revised form 17 June 2025; Accepted 27 June 2025
Available online 16 July 2025
0165-0327/Published by Elsevier B.V.

themselves (e.g., increased mood lability) may be clinically relevant in youth with irritability and are associated with functional impairment (Dejonckheere et al., 2018; Naim et al., 2022; Olthof et al., 2023). EMA can also capture within person change in irritable affect over time rather than only comparing irritability between participants (Verhagen et al., 2016). Thus, EMA may be particularly well suited to increase understanding of the phenomenology of youth irritability (Naim et al., 2021).

As irritability EMA research advances, it is crucial to assess the psychometric validity and reliability of new measures. Research assessing youth irritability via EMA is still nascent (Evans et al., 2023a, 2023b; Flynn et al., 2021; Naim et al., 2021). Past studies have adapted items from previously validated self and parent-reported questionnaires, including the Affective Reactivity Index (Stringaris et al., 2012) and the Positive and Negative Affect Scale (Thompson, 2007), into EMA items. Existing EMA items adapted from the parent- and child-Affective Reactivity Index metrics are necessarily altered to capture momentary affect (Naim et al., 2021). Changes in EMA items and context necessitate additional validation of adapted irritability EMA items to avoid threats to both the replicability and internal validity of irritability EMA studies (Byrne et al., 2023; Flake and Fried, 2020; Mestdagh and Dejonckheere, 2021). Past work has assessed irritability EMA items adopted from the Affective Reactivity Index (Stringaris et al., 2012) in relation to child, parent, and clinician-reported assessments and a Research Domain Criteria (RDoC)-informed behavioral task measuring frustrative non-reward (Naim et al., 2021), as well as items' dynamics (Naim et al., 2022) and network structure (Tseng et al., 2023). In clinically irritable youth, recent work has demonstrated changes in EMA-assessed irritability symptoms during treatment for irritability (Naim et al., in press). However, to date no research has examined the between and within person psychometric properties of irritability EMA items in clinically irritable youth. Psychometric analysis will support the internal validity of irritability EMA measures for future use in clinical research and trials. Importantly, measures must be validated in clinically irritable populations, as well as compared to other clinical and non-clinical youth, to ensure their suitability across symptomatology.

A previous study by Byrne et al. (2023) examined within and between person psychometric properties of anxiety EMA items in anxious and healthy youth, and found evidence for satisfactory psychometrics in youth with anxiety but not in healthy youth, demonstrating the need to compare items across clinical groups. We use data from the same EMA study protocol as Byrne et al. (2023) to examine the between and within person psychometric properties of EMA items capturing irritability in an independent sample of youth with clinically significant irritability who completed EMA measures while completing treatment for irritability (Naim et al., 2024; Naim et al., in press). The current study leveraged the same approach as Byrne et al. (2023), and examined irritability items' between and within person variability, internal consistency, test-retest reliability across two EMA intervals, and convergent and discriminant validity. To facilitate comparison across irritable, anxious, and healthy youth in our sample and that of Byrne et al. (2023), supplemental analyses examined the psychometric properties of anxiety and mood related items in youth with irritability. Consistent with prior work on anxiety EMA items (Byrne et al., 2023), we hypothesized that irritability EMA items would demonstrate acceptable psychometric properties in youth with irritability, including moderate to good test-retest reliability, good convergent and discriminant validity, and sufficient between- and within-person internal consistency as determined by accepted standards for each of our analyses.

2. Methods

2.1. Participants

Participants included 49 youth with chronic irritability ages 8–17 years recruited as part of a larger study protocol in the Washington, D.C. area (Naim et al., 2024; Naim et al., in press). All youth received twelve sessions of exposure-based cognitive behavioral therapy (CBT) for clinically impairing irritability. Sample size was based on availability of data from a completed study treatment cohort (Naim et al., 2024) and an ongoing follow-up treatment study (see Table 1 for demographic information). The sample also overlaps with a recently published study examining changes in EMA-assessed youth irritability symptoms and parent behaviors during CBT treatment for irritability (Naim et al., in press). Full recruitment and screening procedures are outlined in Naim et al. (2024). Eligible youth presented with at least one core symptom of disruptive mood dysregulation disorder (DMDD; i.e., chronically irritable mood and/or temper outbursts), with at least moderate impairment in two or more domains (home, school, and peers) (Naim et al., 2023). While youth frequently presented with multiple diagnoses, primary diagnosis was determined based on youths' chief presenting complaint and clinical judgement of the diagnosis with greatest impairment (Naim et al., 2021) (see Table 1).

2.2. Procedures

All study procedures were approved by the National Institute of Mental Health Institutional Review Board. Participants completed a larger smartphone-based EMA protocol assessing multiple dimensions of irritability, mood, and anxiety symptoms and their situational contexts (Naim et al., 2021). Our study exclusively used iPhones; if participants did not have an iPhone available, participants received a study-provided iPhone. Prior to study participation, research assistants provided a standardized EMA training to ensure familiarity with the study protocol and smartphone, reviewing each EMA item, completing a practice prompt, and ensuring that EMA prompts worked on participants' devices (Naim et al., 2021). EMA prompts used ReTAINE technology (Smith et al., 2019). At each prompt, participants received a text with a

Table 1
Demographic and clinical characteristics of $N = 49$ youth with clinically impairing irritability. Female = sex assigned at birth. Primary Diagnosis = diagnosis of greatest impairment. ARI = Affective Reactivity Index. DMDD = Disruptive Mood Dysregulation Disorder. ODD = Oppositional Defiant Disorder. ADHD = Attention Deficit Hyperactivity Disorder.

	Total sample ($N = 49$)
Female, *n* (%)	18 (36.73)
Age (years), mean (SD)	10.63 (1.83)
Race, *n* (%)	
Asian	1 (2.04)
American Indian or Alaskan Native	0 (0.00)
Black/African American	4 (8.16)
Multiple Races	3 (6.12)
White	38 (77.55)
Not reported	3 (6.12)
Ethnicity, *n* (%)	
Latino or Hispanic	4 (8.16)
Not Latino or Hispanic	42 (85.71)
Not reported	3 (6.12)
Primary diagnosis, n (%)	
DMDD	24 (48.98)
ODD	12 (24.48)
ADHD	13 (26.53)
Interval 1 Questionnaires, mean (sd)	
ARI-S 1 wk Total	4.29 (3.42)
ARI-P 1 wk Total	7.63 (2.67)
Clinician ARI Total	39.38 (14.13)
Interval 2 Questionnaires, mean (sd)	
ARI-S 1 wk Total	4.45 (3.06)
ARI-P 1 wk Total	5.65 (2.61)
Clinician ARI Total	33.87 (15.88)

link to a website delivering the EMA items (see Figs. 1–2). Study-provided iPhones were limited to accessing only the website that delivered the items. Participants had 60 min to respond to each prompt, after which it was coded as missing data. Data was collected directly on a secure server, and no data was saved on study phones. When participants opted to use their own phones, no data was stored on the phone and all surveys were accessed via text message.

Participants completed three seven-day EMA intervals before, during, and following treatment (Naim et al., in press), receiving three prompts daily; we report on the first two intervals in our analyses here. Interval 1 occurred the week prior to their first CBT session, and Interval 2 after session 6 of CBT. To enhance compliance and ease of use, participants selected 60-min time frames during morning (6:00–9:00 AM), afternoon (3:00–6:00 PM), and evening (7:00–10:00 PM) periods to receive each prompt; prompt alerts were then randomized within these windows. Youth were compensated for participation and offered a monetary bonus if they completed 75 % or more of prompts (Naim et al., 2021). Participants were excluded from analyses for each interval if they did not complete at least five prompts in that interval, following prior EMA research (Bylsma et al., 2011; Byrne et al., 2023; Kircanski et al., 2015; Naim et al., 2021). To assess convergent validity with established irritability measures, participants, their parents, and clinicians completed validated irritability assessments (Haller et al., 2020; Stringaris et al., 2012) as closely to each EMA interval as possible.

Two prompts designed to assess irritability symptoms' chronometry (i.e., both pervasive mood states and momentary affect) are the focus of the current analysis. One prompt assessed frustration since the previous prompt to capture irritability throughout the day ("Since the last beep, I felt frustrated"), and another assessed momentary anger at the time of the prompt ("At the time of the beep, I felt annoyed or angry") (Naim et al., 2021). Prompts were broadly modeled on items from the Affective Reactivity Index (e.g., "Angry most of the time") (Stringaris et al., 2012). To assess discriminant validity, analyses included an EMA prompt capturing positive affect, "At the time of the beep, I felt happy." The "Annoyance/Anger At," "Frustration Since", and "Happy" prompts were rated using a 5-point Likert scale, ranging from 1 = Not at all to 5 = Extremely. Of note, participants also completed additional items assessing temper outbursts: e.g., "Since the last beep, I felt really, really angry and out of control" (rated as a categorical "Yes"/"No" response). However, the temper outburst item's endorsement was too low to include in analyses (<7.69 % "Yes" at any prompt in Interval 1). Of the three continuous EMA items, analyses included those allowing for direct comparison with previous psychometric analyses in anxious and healthy youth (Byrne et al., 2023).

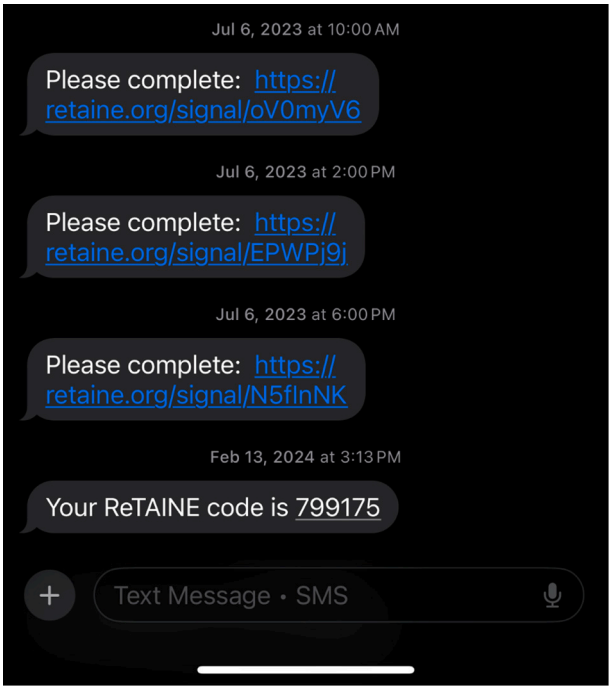


Fig. 1. Convergent and discriminant validity in youth with clinically impairing irritability at Interval 1. SCARED = Screen for Child Anxiety-Related Emotional Disorders; MFQ = Mood and Feelings Questionnaire Total Score Child and Parent Version. *p < 0.05; **p < 0.01.

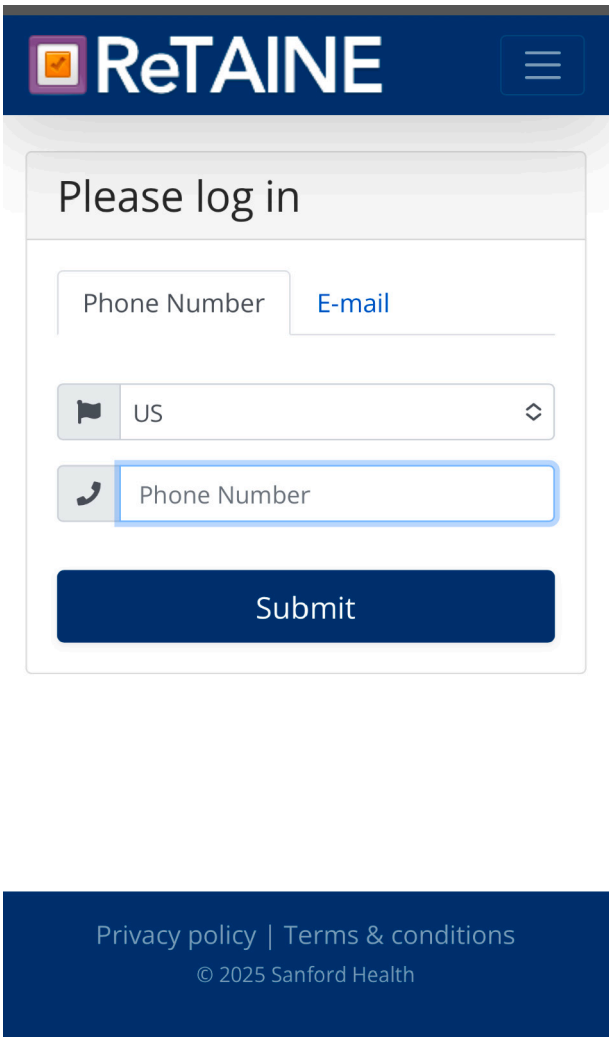


Fig. 2. Login page of Retaine website to complete EMA prompts.

2.3. Materials

2.3.1. Diagnostic status

The KSADS-PL is a semi structured diagnostic interview that assesses present and lifetime DSM-5 disorders (Kaufman et al., 1997). As outlined in Naim et al. (2023), clinicians assessed participants for eligibility and diagnostic categorization using the KSADS-PL, as well as a DMDD supplement (Wiggins et al., 2016).

2.3.2. Child and parent-rated irritability

Irritability symptoms were assessed by youth and their parents via the Affective Reactivity Index Self and Parent 1-Week version as closely to each EMA interval as possible (number of days from completing EMA: Affective Reactivity Index youth-report Mdn = 3.00; Affective Reactivity Index parent-report Mdn = 3.00) (Stringaris et al., 2012). The Affective

Reactivity Index comprises six items pertaining to irritable feelings and behaviors that were summed to generate a total score, and one additional item assessing irritability-related impairment. Each is rated on a 3-point Likert scale from 0 = Not true to 2 = Certainly true. The Affective Reactivity Index has exhibited strong reliability and construct validity (DeSousa et al., 2013; Mulraney et al., 2014; Stringaris et al., 2012). In the current sample, the Affective Reactivity Index-Self exhibited excellent internal consistency in Interval 1 (α = 0.90) and good internal consistency in Interval 2 (α = 0.86). The Affective Reactivity Index-Parent showed good internal consistency for both Intervals 1 (α = 0.85) and 2 (α = 0.80).

2.3.3. Clinician-rated irritability

The Clinician Affectivity Reactivity Index is a 12-item semi-structured interview with good validity and sufficient reliability (Haller et al., 2020). As closely to each EMA interval as possible, clinicians assessed youth irritability via the Clinician Affectivity Reactivity Index (number of days from completing EMA: Mdn = 2.00) (Haller et al., 2020) (see Naim et al., 2023).

2.4. Analyses

Statistical analyses were completed using R version 4.3. Descriptive statistics were computed to assess participant demographics (see Table 1). EMA items' psychometric properties were examined separately by interval, as appropriate. As outlined in Byrne et al. (2023), analyses assessed irritability and positive affect EMA items' variability, internal consistency, test-retest reliability, and convergent and discriminant validity. Within person variability was evaluated by quantifying the degree of change in response from one prompt to the next, and between person variability by assessing the proportion of variability attributed to between subjects differences relative to total variability. Internal consistency was assessed as consistency between both EMA prompts within individuals, and across individuals at each prompt time. Test-retest reliability was evaluated via correlations between mean EMA responses at both timepoints. Convergent validity was assessed by comparing mean EMA responses to validated self, clinician, and parent report measures at each timepoint. Finally, discriminant validity was examined by comparing mean EMA responses to both irritable EMA items with the positive affect EMA item. To facilitate comparison across clinical groups assessed by Byrne et al. (2023) and the current study, supplemental analyses assessed the variability, test-retest reliability, and convergent and discriminant validity of anxiety and mood related EMA measures (see Supplemental Tables 1–2).

2.4.1. Variability

Each participant's within person variability was assessed as moment-to-moment variability in EMA responses at each interval via mean adjusted squared successive differences (MASSD) (Byrne et al., 2023). Unanswered prompts were removed to assess MASSD between all available EMA responses; MASSD accounts for varying lengths between assessments (Funkhouser et al., 2021; Jahng et al., 2008; Trull et al., 2008). To generate a mean MASSD score for each EMA item at each interval, MASSD scores were aggregated across participants (Woyshville et al., 1999). While there are no standardized cutoffs for acceptable variability, higher MASSD scores indicate more within-person variability in participants' responses over time regardless of their mean score, while lower MASSD scores indicate greater stability in responses over time.

Intraclass correlation coefficients (ICC) investigated between person variability in responses of each EMA item at each interval. ICCs were calculated with missing prompt data removed. Here, ICC describes the proportion of an EMA item's inter-subject variability relative to total variability. A higher ICC score indicates more variability in responses between participants (Bolger and Laurenceau, 2013; Snijders and Bosker, 2011). We used the Wald test to evaluate whether the ICC value was significantly greater than zero (Snijders and Bosker, 2011).

2.5. Reliability

2.5.1. Internal consistency

Internal consistency was calculated using a procedure from previous EMA research to account for the multilevel nature of EMA data (Byrne et al., 2023; Viechtbauer, 2017). First, a multivariate multilevel model generated correlations between the two irritability EMA items at the prompt (within each observation) and person (within each participant) level. Prompt and person level correlations excluded missing prompt data. Building on previous approaches (Byrne et al., 2023; Forkmann et al., 2018), internal consistency was calculated (with >0.70 indicating sufficient consistency) between irritability measurements at each observation and within each participant via the Spearman-Brown formula to predict reliabilities:

$$\rho^{*}_{xx'} = \frac{n\rho_{xx'}}{1 + (n-1)\rho_{xx'}},$$

n, number of items; ρ, correlation of items; $\rho_{xx'}$, predicted reliability.

2.5.2. Test-retest reliability

Participants' mean scores were calculated for each EMA item at Intervals 1 and 2. Mean EMA scores for each interval excluded missing prompt values. Each EMA item's test-retest reliability across participants was quantified using the ICC between each item's mean score at both intervals. For these analyses, ICC indicates the correlation between average EMA item responses at each interval. ICC values and 95 % confidence intervals were quantified via a mean-rating ($k = 2$), consistency, two-way mixed effects model. An ICC score < 0.5 indicates poor reliability, an ICC between 0.5 and 0.75 moderate, 0.75–0.9 good, and >0.90 excellent (Koo and Li, 2016; Portney and Watkins, 2009).

2.6. Validity

Convergent validity of the irritability EMA items was assessed via Pearson's r correlations between participants' mean scores and child, parent, and clinician-reported measures of irritability (Affective Reactivity Index-Self; Affective Reactivity Index-P; Clinician Affective Reactivity Index, respectively), with a correlation of <±0.29 considered small, ±0.30 to ±0.49 = moderate, and >±0.50 = strong at the $p < .05$ level. Discriminant validity was evaluated via Pearson's correlations between mean irritability EMA item scores and mean EMA positive affect scores at each interval.

3. Results

The current study initially included $N = 50$ youth receiving CBT treatment for clinically impairing irritability. Demographic data and clinical characteristics are presented in Table 1. One participant was excluded from all EMA analyses due to not responding to at least five EMA prompts in either interval ($N = 49$). At Interval 1, two participants were excluded from analyses due to not responding to at least five EMA prompts ($N = 47$ with sufficient data) but were included in Interval 2. Average compliance rates (percentage of prompts completed) were 80.77 % at Interval 1 and 76.31 % at Interval 2. There were no significant differences between compliance rates between morning, afternoon, and evening prompts in either interval ($ps > .48$). EMA ratings of frustration differed significantly between Interval 1 and Interval 2 ($t(46) = 2.72$, $p < .01$), while other items displayed no significant differences ($ps < .24$). Mean and standard deviation of each EMA item in Interval 1 was calculated by aggregating EMA responses across participants and timepoints (see Table 2). Missing completely at random (MCAR) analyses revealed that EMA data was not missing at random for either Interval 1 ($\chi^2 = 18.73$, $p < .01$) or Interval 2 ($\chi^2 = 21.85$, $p < .01$) (see Discussion).

Table 2
Ecological momentary assessment item variability in youth with clinically impairing irritability at Intervals 1 and 2.

EMA item	Interval	*M*	*SD*	*Min*	*Max*	*M MASSD*	*SD MASSD*	*Min MASSD*	*Max MASSD*	*ICC*
"Since the last beep, I felt frustrated"	Interval 1	1.90	1.21	1.00	5.00	1.31	1.19	0.00	4.67	0.42
	Interval 2	1.70	1.10	1.00	5.00	0.97	1.24	0.00	6.33	0.48
"At the time of the beep, I felt annoyed or angry"	Interval 1	1.66	1.09	1.00	5.00	1.23	1.25	0.00	5.07	0.36
	Interval 2	1.58	1.04	1.00	5.00	0.81	0.89	0.00	3.37	0.43
"At the time of the beep, I felt happy"	Interval 1	3.09	1.24	1.00	5.00	1.77	1.62	0.00	7.26	0.25
	Interval 2	2.99	1.28	1.00	5.00	1.45	1.23	0.00	5.63	0.36

3.1. Within and between person variability

The descriptive statistics and variability of each EMA item at each interval is reported in Table 2. Mean MASSD scores for both irritability EMA items exhibited within person variability at both intervals (MASSD scores ranging 0.81–1.31). A small portion of youth demonstrated no variability within an EMA item with a MASSD score = 0.00 (Interval 1: $n = 8$ "Annoyance/Anger At," $n = 3$ "Frustration Since"; Interval 2: $n = 12$ "Annoyance/Anger At," $n = 9$ "Frustration Since"). The gap between mean and maximum MASSD scores suggests that response variability differed considerably between youth. We observed comparable variability for the "Happy" EMA item at both intervals, demonstrating within person variability across EMA items (1.45–1.77). ICC values demonstrated between-person variability at both intervals for both irritability and "Happy" EMA items (ICC scores ranging 0.25–0.48; see Table 2).

3.2. Reliability

Internal consistency of the two irritability EMA items at both person and prompt levels at each interval and test-retest reliability between Intervals 1 and 2 are presented in Table 3. At each interval, both irritability EMA items exhibited excellent internal consistency at the person level (Interval 1 = 0.97, Interval 2 = 0.91), and lower but acceptable consistency at the prompt level (Interval 1 = 0.62, Interval 2 = 0.53). This is consistent with analyses in anxious and healthy youth populations (Byrne et al., 2023). ICC values indicated good test-retest reliability for irritability EMA measures (*ICC* = 0.84 for the "Frustration Since" prompt and *ICC* = 0.78 for the "Annoyance/Anger At" prompt), and moderate reliability for the positive affect measure (*ICC* = 0.70).

3.3. Validity

Information on convergent and discriminant validity is presented for each EMA item at Interval 1 in Fig. 3. Both irritability EMA items were strongly correlated with one another ($r(45) = 0.90$, $p < .01$) and were moderately correlated with child reported irritability on the Affective Reactivity Index-Self ($r(45) = 0.56$ for the "frustrated" prompt and $r(45) = 0.58$ for the "angry or annoyed" prompt, $ps < .01$). Both EMA items demonstrated small to moderate correlations with clinician reported irritability on the Clinician Affective Reactivity Index ($r(45) = 0.38$ for the "Frustration Since" prompt, $p < .01$ and $r(45) = 0.32$ for the "Annoyance/Anger At" prompt, $p < .05$). Items were not significantly correlated with parent-reported irritability on the Affective Reactivity Index-Parent, with the "Annoyance/Anger At" EMA prompt and Affective Reactivity Index-Parent trending toward significance ($p = .06$). Irritability EMA measures were negatively correlated with the positive affect EMA measure, though this association was not significant, indicating overall good discriminant validity.

Table 3
Internal consistency and test-retest reliability in youth with clinically impairing irritability at Intervals 1 and 2.

	Reliability (person)	Reliability (prompt)
Interval 1 EMA Irritability Items	0.974	0.616
Interval 2 EMA Irritability Items	0.914	0.528

EMA Item	ICC	ICC 95 % CI
"Since the last beep, I felt frustrated"	0.836	0.723 to 0.905
"At the time of the beep, I felt annoyed or angry"	0.782	0.64 to 0.873
"At the time of the beep, I felt happy"	0.699	0.518 to 0.821

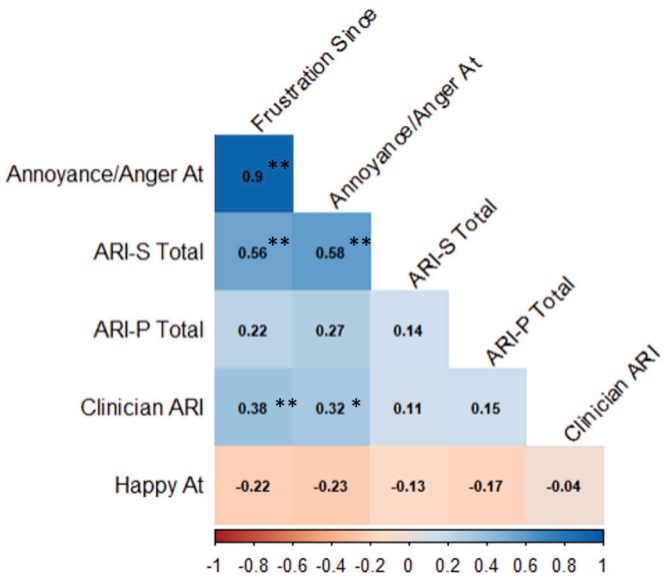


Fig. 3. Convergent and discriminant validity in youth with clinically impairing irritability at Interval 1. *p < .05; **p < .01.

4. Discussion

We examined the psychometric properties of irritability EMA items in a sample of treatment-seeking youth with irritability. Critically, the irritability EMA items demonstrated acceptable variability, consistency, reliability, and both convergent as well as discriminant validity in youth with irritability, and thus may be suitable for clinical populations and trials. These findings also extend previous work demonstrating convergence between these irritability EMA items and existing clinical and behavioral measures, facilitating studying irritability at multiple levels of analysis to refine conceptions of youth irritability and ultimately inform treatment (Joyner and Perkins, 2023; Naim et al., 2021). Below, these findings are discussed in context with psychometric analysis of EMA items by Byrne et al. (2023) in anxious and healthy youth.

As hypothesized, the EMA irritability items demonstrated within and between person variability. While there is no agreed-upon standard for variability in EMA research, results were similar to those reported by Byrne et al. (2023) and suggest that these items capture differences both within individual response patterns and between participants. This extends findings reported by Byrne et al. (2023) that anxiety EMA items in this study protocol sufficiently capture between and within person variability in anxious youth. In supplemental analyses, Byrne et al. (2023) examined irritability EMA items from this protocol in anxious

and healthy youth. They found evidence for within and between person variability in youth with anxiety, but weaker within person variability in healthy youth, suggesting that these items may not sufficiently capture irritability fluctuations in healthy populations. These results are not totally unexpected given findings by (Naim et al., 2022) that youth with DMDD demonstrate significantly more overall negative affect lability compared to healthy youth; healthy youth may thus demonstrate fewer overall fluctuations in irritable affect. Therefore, while these irritability items may be especially useful to capture irritability symptoms within clinical trials, additional measures or changes may be necessary to be adequately sensitive to dynamic fluctuations in irritability in healthy comparison populations (Byrne et al., 2023). Additionally, similar to Byrne et al. (2023), the "Happy" EMA item exhibited within and between person variability in this clinically irritable sample and may be suited for future research on positive and negative affect in youth with irritability (Naim et al., 2022; Vogel et al., 2023).

The irritability EMA items exhibited excellent internal consistency at the person level, and lower but sufficient consistency at the prompt level. These findings were consistent with hypotheses, and parallel to findings by Byrne et al. (2023) with anxiety EMA items in anxious youth in this protocol. Indeed, as Byrne et al. (2023) note, lower prompt-level reliability may reveal meaningful fluctuations in affect across time and contexts (Mischel et al., 2002). Unlike approaches that aggregate EMA scores across timepoints, variation in prompt-level responses reflects both pervasive mood states and momentary affect. Such measures may be clinically meaningful in irritable youth and vary by symptom presentation (Dejonckheere et al., 2018; Naim et al., 2022; Olthof et al., 2023). These findings underscore these EMA items' utility for clinical research. Further, despite involvement in CBT treatment and a significant change in EMA responses for the "frustration" prompt in these analyses, both EMA items exhibited good test-retest reliability across intervals, indicating they are measures of stable constructs.

EMA irritability items demonstrated good convergent validity. Both irritability items were strongly correlated with child-reported irritability and moderately correlated with clinician-rated irritability, in line with previous evidence of convergence between irritability EMA items and youth and clinician-reported irritability (Naim et al., 2021). However, items were not significantly correlated with parent-reported irritability. This contrasts with findings by Byrne et al. (2023) of a small correlation between EMA anxiety items and parent reported anxiety in anxious youth. However, current findings align with previous research demonstrating marked informant discrepancies between youth and parent reported irritability; particular clinical and demographic features may contribute to discrepancies (see Mallidi et al., 2023), which may also reflect different underlying irritability-related constructs (De Los Reyes et al., 2009; Zik et al., 2022). The lack of convergence between EMA irritability items and parent-reported irritability also diverges from Naim et al. (2021), whose findings demonstrate convergence between irritability EMA and parent-reported irritability in this protocol. However, while their analyses used a transdiagnostic sample, the current study's focus on youth with clinical irritability undergoing CBT treatment may underscore the particular importance of obtaining multiple informant reports of youth irritability in clinical contexts. Finally, both irritability EMA items exhibited a negative but non-significant correlation with the positive affect EMA item, while they were highly correlated with one another ($r = 0.90$), indicating discriminant validity.

Our study has several potential implications for future research in clinically irritable youth. Irritability may involve more dynamic affect changes than other emotional symptoms (e.g., outbursts) (Naim et al., 2022). Future work could use robust EMA measures of irritability to understand dynamic changes in irritable affect, as well as influence of contextual factors on irritability symptoms (e.g., parental or peer interactions) (Naim et al., 2021). Further, EMA assessment of real time irritability symptoms during treatment may facilitate understanding of treatment response and tailoring of treatment to individual youths' needs (Naim et al., 2021). Finally, examining changes in irritable affect over the course of treatment, such as the significant decrease in reported frustration in the current sample, may allow researchers to test hypotheses on putative mechanisms of treatment efficacy (Brotman et al., 2017).

Despite demonstration of the robustness of the psychometric properties of these irritability EMA items, this study had several important limitations. First, participants' involvement in CBT treatment during the EMA intervals creates a potential confound in the interpretation of results, particularly reliability analyses, though good test-retest reliability was observed despite evidence of response change in one item. Second, similar to findings from Byrne et al. (2023) that anxiety-related EMA items from this study protocol were valid in participants with anxiety but not healthy volunteers, supplemental analyses demonstrated limited within person variability in for these anxiety items in participants with primary irritability, despite frequent clinical co-occurrence of irritability and anxiety symptoms (Stoddard et al., 2014). However, anxious EMA items significantly correlated with both parent and child reported anxiety in youth with irritability; diminished within person anxiety may thus reflect reduced fluctuations in anxiety in the current sample. Third, while analyses suggest these EMA items' ability to capture dynamic fluctuations in affect, the binary nature and low endorsement of the outburst related EMA item prevented its inclusion in these analyses. These EMA items thus do not capture outbursts, which are highly salient clinically and a target of novel interventions (Naim et al., 2023). Future work should aim to combine these items with EMA measures better able to capture phasic dimensions of irritability, especially for use in clinical trials. Fourth, though overall EMA compliance was high, the EMA data in each study interval were revealed to be missing not at random. Though all included participants met the accepted cutoff for data completion, a few outlier participants with relatively low compliance rates may have driven this result. Additionally, while the study assessed child, parent, and clinician reports of irritability as closely as possible to each EMA interval, these did not always exactly map on to the EMA interval being studied, which may have impacted findings.

Lastly, the limitations of our particular sample should be considered. The accessibility of outpatient treatment at the National Institute of Mental Health (NIMH) and the unique nature of the clinical irritability phenotype may have impacted the generalizability of the current study. Specifically, our study had a relatively small sample size, was majority white and male, and all youth and parents were receiving treatment at the NIMH. Our sample had the time and ability to enroll in a time intensive treatment study and complete additional measures for research, which may create a selection bias relative to the general population (e.g., higher compliance rates). Finally, we provided free study phones and standardized training on EMA measures that may not be easily replicated in a typical clinical setting. Future research should replicate these findings using larger samples across diverse populations.

5. Conclusions

Analyses demonstrate that EMA items designed to capture irritability exhibit acceptable between and within person psychometric properties, including between and within person variability, internal consistency, test-retest reliability, and convergent and discriminant validity in youth with irritability, suggesting items are psychometrically sound and fit for use with clinically irritable populations in clinical trials. Results align with previous research demonstrating strong psychometric properties of anxiety EMA items in anxious participants (Byrne et al., 2023), as well as the need for comparison across clinical groups. As EMA increases in popularity in youth psychopathology research, it is essential to validate emerging measures to avoid threats to replication and internal validity. This study uniquely evaluates the psychometric properties of irritability EMA items while accounting for the dynamic nature of EMA data, facilitating their use in future clinical research.

There are no acknowledgements the authors wish to disclose.

CRediT authorship contribution statement

Eleanor G. Hansen: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis. **Reut Naim:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Lauren M. Henry:** Writing – review & editing. **Katharina Kircanski:** Writing – review & editing, Methodology, Conceptualization. **Daniel S. Pine:** Writing – review & editing. **Melissa A. Brotman:** Writing – review & editing, Methodology, Conceptualization. **Meghan E. Byrne:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization.

Funding statement

This work was supported by the Intramural Research Program (IRP) of the National Institute of Mental Health/National Institutes of Health (NIMH/NIH), ZIAMH002781 (Pine), ZIAMH002786 (Leibenluft), ZIAMH002778 (Leibenluft), and conducted under NIH Clinical Study Protocols 01-M-0192 and 02-M-0021 (ClinicalTrials.gov identifiers: NCT00018057 and NCT00025935).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jad.2025.119788.

References

Bolger, N., Laurenceau, J.-P., 2013. Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research. Guilford press.
Brotman, M.A., Kircanski, K., Stringaris, A., Pine, D.S., Leibenluft, E., 2017. Irritability in youths: a translational model. Am. J. Psychiatry 174 (6), 520–532.
Bylsma, L.M., Taylor-Clift, A., Rottenberg, J., 2011. Emotional reactivity to daily events in major and minor depression. J. Abnorm. Psychol. 120 (1), 155.
Byrne, M.E., Bernstein, R.A., Pine, D.S., Kircanski, K., 2023. Ecological momentary assessment of youth anxiety: evaluation of psychometrics for use in clinical trials. J. Child Adolesc. Psychopharmacol. 33 (10), 409–417.
Collishaw, S., Maughan, B., Natarajan, L., Pickles, A., 2010. Trends in adolescent emotional problems in England: a comparison of two national cohorts twenty years apart. J. Child Psychol. Psychiatry 51 (8), 885–894.
De Los Reyes, A., Henry, D.B., Tolan, P.H., Wakschlag, L.S., 2009. Linking informant discrepancies to observed variations in young children's disruptive behavior. J. Abnorm. Child Psychol. 37 (5), 637–652.
Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., Brose, A., Bastian, B., Kuppens, P., 2018. The bipolarity of affect and depressive symptoms. J. Pers. Soc. Psychol. 114 (2), 323.
DeSousa, D.A., Stringaris, A., Leibenluft, E., Koller, S.H., Manfro, G.G., Salum, G.A., 2013. Cross-cultural adaptation and preliminary psychometric properties of the Affective Reactivity Index in Brazilian Youth: implications for DSM-5 measured irritability. Trends Psychiatry Psychother. 35, 171–180.
Evans, S.C., Shaughnessy, S., Karlovich, A.R., 2023a. Future directions in youth irritability research. Journal of Clinical Child & Adolescent Psychology 52 (5), 716–734.
Evans, S.C., Corteselli, K.A., Edelman, A., Scott, H., Weisz, J.R., 2023b. Is irritability a top problem in youth mental health care? A multi-informant, multi-method investigation. Child Psychiatry Hum. Dev. 54 (4), 1027–1041.
Flake, J.K., Fried, E.I., 2020. Measurement schmeasurement: questionable measurement practices and how to avoid them. Adv. Methods Pract. Psychol. Sci. 3 (4), 456–465.
Flynn, M.M., Rosen, P.J., Reese, J.S., Slaughter, K.E., Alacha, H.F., Olczyk, A.R., 2021. Examining the influence of irritability and ADHD on domains of parenting stress. Eur. Child Adolesc. Psychiatry 1–14.
Forkmann, T., Spangenberg, L., Rath, D., Hallensleben, N., Hegerl, U., Kersting, A., Glaesmer, H., 2018. Assessing suicidality in real time: a psychometric evaluation of self-report items for the assessment of suicidal ideation and its proximal risk factors using ecological momentary assessments. J. Abnorm. Psychol. 127 (8), 758.
Funkhouser, C.J., Kaiser, A.J., Alqueza, K.L., Carrillo, V.L., Hoffman, L.M., Nabb, C.B., Auerbach, R.P., Shankman, S.A., 2021. Depression risk factors and affect dynamics: an experience sampling study. J. Psychiatr. Res. 135, 68–75.
Haller, S.P., Kircanski, K., Stringaris, A., Clayton, M., Bui, H., Agorsor, C., Cardenas, S.I., Towbin, K.E., Pine, D.S., Leibenluft, E., 2020. The clinician affective reactivity index: validity and reliability of a clinician-rated assessment of irritability. Behav. Ther. 51 (2), 283–293.
Henry, L.M., Hansen, E., Chimoff, J., Pokstis, K., Kiderman, M., Naim, R., Kossowsky, J., Byrne, M.E., Lopez-Guzman, S., Kircanski, K., 2024. Selecting an ecological momentary assessment platform: tutorial for researchers. J. Med. Internet Res. 26, e51125.
Jahng, S., Wood, P.K., Trull, T.J., 2008. Analysis of affective instability in ecological momentary assessment: indices using successive difference and group comparison via multilevel modeling. Psychol. Methods 13 (4), 354.
Joyner, K.J., Perkins, E.R., 2023. Challenges and Ways Forward in Bridging Units of Analysis in Clinical Psychological Science.
Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., Williamson, D., Ryan, N., 1997. Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): initial reliability and validity data. J. Am. Acad. Child Adolesc. Psychiatry 36 (7), 980–988.
Kircanski, K., Thompson, R.J., Sorenson, J.E., Sherdell, L., Gotlib, I.H., 2015. Rumination and worry in daily life: examining the naturalistic validity of theoretical constructs. Clin. Psychol. Sci. 3 (6), 926–939.
Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15 (2), 155–163.
Leibenluft, E., 2017. Pediatric irritability: a systems neuroscience approach. Trends Cogn. Sci. 21 (4), 277–289.
Mallidi, A., Meza-Cervera, T., Kircanski, K., Stringaris, A., Brotman, M.A., Pine, D.S., Leibenluft, E., Linke, J.O., 2023. Robust caregiver-youth discrepancies in irritability ratings on the affective reactivity index: an investigation of its origins. J. Affect. Disord. 332, 185–193.
Mestdagh, M., Dejonckheere, E., 2021. Ambulatory assessment in psychopathology research: current achievements and future ambitions. Curr. Opin. Psychol. 41, 1–8.
Mischel, W., Shoda, Y., Mendoza-Denton, R., 2002. Situation-behavior profiles as a locus of consistency in personality. Curr. Dir. Psychol. Sci. 11 (2), 50–54.
Mulraney, M.A., Melvin, G.A., Tonge, B.J., 2014. Psychometric properties of the affective reactivity index in Australian adults and adolescents. Psychol. Assess. 26 (1), 148.
Naim, R., Smith, A., Chue, A., Grassie, H., Linke, J., Dombek, K., Shaughnessy, S., McNeil, C., Cardinale, E., Agorsor, C., 2021. Using ecological momentary assessment to enhance irritability phenotyping in a transdiagnostic sample of youth. Dev. Psychopathol. 33 (5), 1734–1746.
Naim, R., Shaughnessy, S., Smith, A., Karalunas, S.L., Kircanski, K., Brotman, M.A., 2022. Real-time assessment of positive and negative affective fluctuations and mood lability in a transdiagnostic sample of youth. Depress. Anxiety 39 (12), 870–880.
Naim, R., Dombek, K., German, R.E., Haller, S.P., Kircanski, K., Brotman, M.A., 2023. An exposure-based cognitive–behavioral therapy for youth with severe irritability: feasibility and preliminary efficacy. Journal of Clinical Child & Adolescent Psychology 1–17.
Naim, R., Dombek, K., German, R.E., Haller, S.P., Kircanski, K., Brotman, M.A., 2024. An exposure-based cognitive–behavioral therapy for youth with severe irritability: feasibility and preliminary efficacy. J. Clin. Child Adolesc. Psychol. 53 (2), 260–276.
Naim, R., Pandya, U., Shaughnessy, S., German, R.E., Henry, L.M., Kircanski, K., Brotman, M.A., 2025. Advancing the Measurement of Psychotherapy Outcomes for Youth with Irritability Using In-Vivo Ecological Momentary Assessment (Manuscript in press).
Olthof, M., Hasselman, F., Aas, B., Lamoth, D., Scholz, S., Daniels-Wredenhagen, N., Goldbeck, F., Weinans, E., Strunk, G., Schiepek, G., 2023. The best of both worlds? General principles of psychopathology in personalized assessment. Journal of psychopathology and clinical science 132 (7), 808.
Portney, L.G., Watkins, M.P., 2009. Foundations of Clinical Research: Applications to Practice, 892. Pearson, Prentice Hall Upper Saddle River, NJ.
Shiffman, S., Stone, A.A., Hufford, M.R., 2008. Ecological momentary assessment. Annu. Rev. Clin. Psychol. 4, 1–32.
Smith, A.R., Kircanski, K., Brotman, M.A., Do, Q.B., Subar, A.R., Silk, J.S., Engel, S., Crosby, R.D., Harrewijn, A., White, L.K., 2019. Advancing clinical neuroscience through enhanced tools: pediatric social anxiety as an example. Depress. Anxiety 36 (8), 701–711.
Snijders, T.A., Bosker, R., 2011. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.
Stoddard, J., Stringaris, A., Brotman, M.A., Montville, D., Pine, D.S., Leibenluft, E., 2014. Irritability in child and adolescent anxiety disorders. Depress. Anxiety 31 (7), 566–573.
Stringaris, A., Cohen, P., Pine, D.S., Leibenluft, E., 2009. Adult outcomes of youth irritability: a 20-year prospective community-based study. Am. J. Psychiatry 166 (9), 1048–1054.
Stringaris, A., Goodman, R., Ferdinando, S., Razdan, V., Muhrer, E., Leibenluft, E., Brotman, M.A., 2012. The Affective Reactivity Index: a concise irritability scale for clinical and research settings. J. Child Psychol. Psychiatry 53 (11), 1109–1117.
Thompson, E.R., 2007. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). J. Cross Cult. Psychol. 38 (2), 227–242.
Trull, T.J., Ebner-Priemer, U., 2013. Ambulatory assessment. Annu. Rev. Clin. Psychol. 9, 151–176.
Trull, T.J., Solhan, M.B., Tragesser, S.L., Jahng, S., Wood, P.K., Piasecki, T.M., Watson, D., 2008. Affective instability: measuring a core feature of borderline personality disorder with ecological momentary assessment. J. Abnorm. Psychol. 117 (3), 647.
Tseng, W.L., Naim, R., Chue, A., Shaughnessy, S., Meigs, J., Pine, D.S., Leibenluft, E., Kircanski, K., Brotman, M.A., 2023. Network analysis of ecological momentary assessment identifies frustration as a central node in irritability. J. Child Psychol. Psychiatry 64 (8), 1212–1221.

Verhagen, S.J., Hasmi, L., Drukker, M., van Os, J., Delespaul, P.A., 2016. Use of the experience sampling method in the context of clinical trials. BMJ Ment Health 19 (3), 86–89.
Viechtbauer, W., 2017. Reliability of ESM assessments of mood and mood sensitivity. In: Digital health in ambulatory assessment–Abstract book of the 5th Biennial Conference of the Society for Ambulatory Assessment.
Vogel, A.C., Brotman, M.A., Roy, A.K., Perlman, S.B., 2023. Defining positive emotion dysregulation: integrating temperamental and clinical perspectives. J. Am. Acad. Child Adolesc. Psychiatry 62 (3), 297–305.
Wiggins, J.L., Brotman, M.A., Adleman, N.E., Kim, P., Oakes, A.H., Reynolds, R.C., Chen, G., Pine, D.S., Leibenluft, E., 2016. Neural correlates of irritability in disruptive mood dysregulation and bipolar disorders. Am. J. Psychiatry 173 (7), 722–730.
Woyshville, M.J., Lackamp, J.M., Eisengart, J.A., Gilliland, J.A., 1999. On the meaning and measurement of affective instability: clues from chaos theory. Biol. Psychiatry 45 (3), 261–269.
Zik, J., Deveney, C.M., Ellingson, J.M., Haller, S.P., Kircanski, K., Cardinale, E.M., Brotman, M.A., Stoddard, J., 2022. Understanding irritability in relation to anger, aggression, and informant in a pediatric clinical population. J. Am. Acad. Child Adolesc. Psychiatry 61 (5), 711–720.