UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Cloos, L., Siepe, B. S., Piccirillo, M. L., Fried, E., Wang, S. B., Helmich, M. A., Johnson, S. U., Hoffart, A., & Ebrahimi, O. V. (2025). Accuracy and Consistency of Visual Analog Scales in Ecological Momentary Assessment and Digital Studies. *Collabra: Psychology*, *11*(1). https://doi.org/10.1525/collabra.142735

Methodology and Research Practice

Accuracy and Consistency of Visual Analog Scales in Ecological Momentary Assessment and Digital Studies

Leonie Cloos[1][a], Björn S. Siepe[2], Marilyn L. Piccirillo[3], Eiko Fried[4], Shirley B. Wang[5], Marieke A. Helmich[6], Sverre Urnes Johnson[6,7], Asle Hoffart[8], Omid V. Ebrahimi[7,9][b]

[1] Department of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium, [2] Department of Psychology, Philipps University of Marburg, Marburg, Germany, [3] Department of Psychiatry, Rutgers, The State University of New Jersey, Piscataway, NJ, USA, [4] Department of Clinical Psychology , Leiden University, Leiden, The Netherlands, [5] Department of Psychology, Yale University, New Haven, CT, USA, [6] Department of Psychology, University of Oslo, Oslo, Norway, [7] Psychiatric Hospital and Research Center, Modum Bad – Gordon Johnsens stiftelse, Vikersund, Norway, [8] Psychiatric Hospital and Research Center, Modum Bad – Gordon Johnsens stiftelse, Vikersund, Norway, [9] Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom

Keywords: Measurement accuracy, Scale interpretation, Reactivity, Response shifts, Self-report, Digital assessment, Ecological Momentary Assessment (EMA)

https://doi.org/10.1525/collabra.142735

Collabra: Psychology

Vol. 11, Issue 1, 2025

The ubiquity of digital technologies has increased assessments of thoughts, behaviors, and experiences via electronic devices. Surveys on smartphones or laptops often implement Visual Analogue Scales (VAS), recording responses on a continuous slider (0-100). This is particularly relevant for data collection in daily life, such as ecological momentary assessments (EMA), which repeatedly present items on mobile devices. However, the accuracy of digital VAS has been questioned, particularly regarding tactile precision (e.g., ability to accurately select values) and the consistency of scale interpretation both between- and within-persons over time (e.g., change in scale interpretation or reactivity to repeated measures). Participants (N = 3,761, 67.03% female; M_{age} = 47.09; SD = 14.41) from the Critical Incidents and Psychological Adaptation (CIPA) Study completed a 30-day EMA assessment. We investigated the accuracy of VAS in terms of (1) tactile precision, (2) respondents' perception of the neutral point post-EMA, and (3) test-retest consistency of affect ratings pre- and post-EMA. (1) Tactile precision was assessed by asking participants to enter exactly 31 on a 0-100 slider. Results showed high precision (M = 31.01; SD = 3.28; 87.0% scored between 30-32). (2) Between-person agreement on scale perception was assessed by asking participants to determine the neutral score on two affect items (unipolar and bipolar). 82.19% and 88.89% indicated the expected scale midpoint (50 and 0, ± 5) as neutral, respectively. Neutral points deviating from the expected midpoint were correlated (r = .71-.73) with the person-specific means across the EMA period on the respective item. (3) Test-retest consistency was evaluated by asking participants to rate how happy/sad they/others would rate affective events (e.g., a serious argument) pre- and post-EMA. Consistency across time was high (*median change* = 0-5). Findings support the accuracy and consistency of digital VAS, within the scope of the current methods.

The ownership of digital devices across the population has allowed for the assessment of individuals' feelings, thoughts, and behaviors in daily life. Although there is a rapid development of different smartphone applications or large survey tools to collect digital data, there is little methodological and psychometric research to inform the quality of the data collected with digital devices (Chmielewski & Kucker, 2020; Van Berkel et al., 2020).

Modern computerized administration of questionnaires on electronic (mobile) devices allows for a convenient implementation of the Visual Analog Scale (VAS), where respondents rate questions on a continuum rather than choosing between categories that are typical for Likert-scales (García-Pérez & Alcalá-Quintana, 2023). VAS ratings were developed to address the limitations of categorical scales by facilitating assessment of continuous phenomena, such as feelings, to be measured without artificial cate-

a Leonie Cloos and Björn Siepe share the first authorship.

b Correspondence concerning this article should be addressed to Omid V. Ebrahimi: omid.ebrahimi@psy.ox.ac.uk

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

gories (Warriner et al., 2017). Typically, VAS scales range from 0 to 100 (unipolar scale), a range derived from paper-and-pencil studies that found the 100 mm line easy to score with a ruler (Aitken, 1969; Yeung & Wong, 2019). Bipolar scales are also commonly employed, where the points are distributed from -50 to 50, with a neutral mid-point of 0, to measure positive and negative aspects (or liking/disliking) of a construct (Setnik et al., 2017). Compared to ordinal scales, VAS allow for greater differentiation and engage participants to adjust their ratings to greater precision, thereby providing a more fine-grained and accurate picture of the construct being measured (Funke & Reips, 2012). The graphical interface of mobile devices enhances the usability and appeal of VAS for surveys, making them a frequently used response format (Reips & Funke, 2008). Unlike traditional paper-and-pencil methods that require scoring with a ruler, data-collection platforms automatically store responses based on the slider's position. For example, if a user moves the slider to 61% of the scale's range, the system directly records the score as 61.

Researchers in the social, health, and organizational sciences are increasingly collecting data via electronic (mobile) devices, for example, using ecological momentary assessment (EMA; Junghaenel & Stone, 2020; Myin-Germeys & Kuppens, 2022) or digital diaries (Gunthert & Wenze, 2012; Saltzman et al., 2021). Despite their widespread use, there is a lack of studies assessing the psychometric properties of electronically administered VAS (Abend et al., 2014). This is especially challenging because VAS are often used as single-item measures, leading to concerns about their accuracy and consistency (Allen et al., 2022). These concerns include (1) tactile precision (i.e., selecting the intended rating on the slider), (2) between-person consistency in scale interpretation (i.e., extent to which the interpretation of the midpoints of scales is consistent across participants), and (3) test-retest consistency (i.e., providing the same rating for the same item across different instances or time points) related to concerns about response shifts.

These issues are frequently discussed among EMA researchers (see Fritz et al., 2023; Stinson et al., 2022; Stone et al., 2023), as concerns about accuracy and consistency of smartphone assessments are particularly relevant in daily life studies, which increasingly rely on smartphone applications to collect real-time self-reports (Mestdagh & Dejonckheere, 2021). Some researchers have compared VAS and Likert scales with much debate; some evidence and recommendations point towards the use of VAS measures for EMA contexts (Haslbeck et al., 2023, 2025). However, the comparison of VAS and Likert scale has limited value without first establishing a clearer understanding of the accuracy and precision of the VAS structure first. Thus, we aim to address the above-mentioned concerns about VAS accuracy and precision with empirical data.

The first issue addressed in this paper is that of (1) tactile precision, defined as the ability to accurately select the intended response on the VAS (e.g., '81' on a 0-100 slider). This concern is particularly relevant for VAS sliders, as their response ranges are extensive and the numerous response points lack annotations, making precise selection challenging (García-Pérez & Alcalá-Quintana, 2023). The problem is further exacerbated on smaller smartphone screens, where distinctions between response options become limited due to a decrease in the size of the slider (van Berkel, 2017). Another issue with larger screen sizes is that certain areas of the screen become less accessible for one-handed use (e.g., with the thumb), making responses more effortful (Karlson et al., 2006). Similarly, older participants (Wenz & Keusch, 2023) may have a harder time handling smartphones and scoring precisely. In this study, we (1) investigate whether participants can precisely hit a score on the rating scale and if demographic (e.g., age or gender) and physical characteristics (e.g., Body Mass Index; BMI) are related to deviations from this precision.

A second source of imprecision may arise from (2) the between-person differences in how individuals interpret the range of the VAS slider and define key points, such as the neutral point on the scale. That is, VAS measures may not accurately capture information due to between-person differences in how individuals perceive and interpret a continuous scale, particularly when evaluating complex subjective experiences like affect. Arguably, this problem is less pronounced in 5- or 7-point Likert scales (Simms et al., 2019) where each scale point can have a specific meaning (e.g., 1 = not at all, 2 = very rarely, 3 = occasionally, etc.). A recent meta-study comparing Likert and VAS ratings in EMA studies revealed that VAS data often exhibit bimodal distributions (Haslbeck et al., 2023). One possible explanation for bimodal or multimodal distributions in VAS ratings could be participants anchoring their responses differently, particularly around their subjective interpretation of a neutral point. If participants differ in their definition of the location of the neutral point, it can complicate the comparison of ratings between individuals. For instance, when asked to rate their current stress, some participants may interpret a "0" as meaning "not at all stressed," while others might see "0" as a neutral or baseline level. Furthermore, the interpretation of the researcher and the participants may differ, where participants may interpret the midpoint as their personal 'baseline', which can be on average quite happy, whereas researchers assume that participants truly 'feel neutral'. It is therefore important to understand where participants determine the neutral point on a VAS slider.

To investigate this further in the context of EMA studies, we explored (2) how participants define their "neutral" point by asking them to identify this on two commonly used types of VAS scales (i.e., unipolar and bipolar). During the EMA study, participants rated their affect on two single items measuring bipolar affect (negative to positive) and unipolar affect (only the positive dimension); on the post-EMA survey they were asked to identify the score they used as neutral reference point. This approach aimed to better understand individual differences in scale interpretation and their potential impact on self-reports.

A third source of concern regarding imprecision is (3) a response shift in the interpretation of the response scale over time. Test-retest consistency refers to an individual's consistency in their response as a function of completing repeated assessments (e.g., whether an 80 out of 100 rep-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

resents the same level of happiness for a participant on day 1 of the study vs. on day 30). After the period of repeated assessment, participants' conceptualization of the measurement target may shift due to changes in their priorities, evolving definitions of the phenomenon being measured, or variations in experiences (Schwartz et al., 2004). In EMA studies, this is particularly relevant as repeated exposure to the items may increase response shift bias (e.g., due to reactivity or desensitization), where participants' initial responses differ systematically from later responses (König et al., 2022; Shrout et al., 2018). The prevalence of response shifts, including initial response elevation bias in EMA data, and the factors influencing these shifts are not yet fully understood (Anvari et al., 2023; Arslan et al., 2020; Cerino et al., 2022; Eisele et al., 2023). To investigate this, we examined potential changes in response scale interpretation or reactivity to an EMA study by investigating the (3) test-retest consistency of participants' ratings of affect levels related to different scenarios before and after the EMA period.

In this study, we examined three concerns regarding the accuracy and consistency of VAS measures in the context of an EMA study with 3,761 participants. These concerns include (1) tactile precision, (2) between-person scale agreement, and (3) test-retest consistency or potential response shifts in affective ratings before and after EMA measurement. These analyses provide insights into the accuracy and consistency of VAS ratings in mobile device-based studies.

Methods

Study Design and Participants

This study is part of a large-scale nationwide study of the Norwegian adult population which commenced in March 2023 (Critical Incidents and Psychological Adaptation; The CIPA Study: https://www.cipastudy.com). Adults were randomly sampled from each region of Norway using population registries to ensure a geographically representative sample and were invited to the study via email. The study received ethical approval from the Regional Committee for Medical and Health Research Ethics (reference: 522020). The prospective study follows over 20,000 participants with longitudinal panel measures over a 15-year period. Part of the sample (N = 3,761) was assessed with an EMA design, collecting four daily measurements over a 30-day period, with the EMA data used in the present study.

Procedures

The 3,761 adults registered for the EMA study received four daily surveys to assess daily patterns in affect, cognition, and mental health for 30 consecutive days (t_{max} = 120 EMA assessments per person). The daily assessments were spaced equidistantly, with 4 hours in between each assessment. The assessments were programmed using the smartphone application m-Path (https://m-path.io), which sent the surveys to participants' smartphones with a notification to complete them. Participants had 90 minutes to respond to each survey before the assessment expired.

Before the EMA assessment, participants filled in a baseline survey assessing demographic characteristics and a pre-EMA assessment including 16 affective scenarios. After the 30-day EMA period, participants received a post-EMA survey that assessed their interactions with the EMA assessments over the course of the study. The survey included items on the tactile precision, assessed between-person consistency about the neutral point on VAS scales, and measured the same 16 affective scenarios from the pre-EMA assessment for comparison. Participants had the opportunity to be randomly drawn to receive gift cards as reward for their participation. Among the 3,761 who registered for the EMA study and responded to the pre-EMA assessment, 2,718 (72.26%) participants completed the post-EMA survey.

Measurement

Demographic Characteristics

Demographic characteristics were assessed in the baseline survey, including age, biological sex, education level, employment status, and current self-reported psychiatric diagnosis.

Tactile Precision

In the post-EMA survey, tactile precision was measured using the following item:

> "You will now be asked to attempt to enter a specific number in **one** attempt. **One** attempt means that you can only press and hold the slider marker once before placing it, without lifting your hands and without trying again. You can drag the slider back and forth while holding it down. However, you cannot try again once you have released the slider. Please try your best to provide the number "31" on the first attempt."

Between-Person Consistency in Perceptions of the Neutral Point

Between-person consistency in scale interpretation about what participants perceived as "neutral" responses was assessed using two items included in the post-EMA survey. The first item included the phrasing "I feel happy" with a unipolar VAS response scale that ranged from 0 (*Not at all*) to 100 (*Extremely*) scale. The second item included the item prompt, "I feel …" and used a bipolar VAS response scale that ranged from -50 (*Very bad*) to 50 (*Very good*). For both items, participants were instructed to select the score that they believed would indicate 'feeling neutral' in their VAS ratings.

Test-Retest Consistency

Test-retest consistency was investigated using a series of 16 affective scenarios that were given to participants at the start and the end of the EMA study (i.e., approximately 30 days apart). Participants were presented with four sit-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

uations that varied in valence and intensity: 1) Having a serious argument with somebody close to you; 2) Losing your wallet; 3) Being on a very good vacation; 4) Spending time with good friends. Participants then completed four items regarding how a) happy and b) sad they believed c) they themselves and d) others would feel in each of the four situations. The items "How happy would you/others feel?" and "How sad would you/others feel?" were rated on a scale, with "Not at all" shown on one end (i.e., left side) of the VAS scale and "Extremely" on the right side of the scale. Following evidence that initialization of a slider at the midpoint might induce multimodality (Haslbeck et al., 2023), a line without any starting value was shown to participants. During the EMA period, the same affective items (how happy and how sad participants felt in the moment) were asked at each EMA survey (up to 120 times). This repeated exposure may lead to a response shift, in which the post-EMA responses to the scenarios may differ from the pre-EMA responses. The test-retest consistency measures therefore enable an examination of whether response shifts in scale interpretation have occurred from the pre- to post-EMA survey after repeated exposure to these affective items during the EMA protocol.

Data Analysis

We operationalized (1) tactile precision as the absolute difference from the target response of '31' and investigated the descriptive statistics of tactile imprecision across participants. We then tested the associations between tactile precision and a range of demographic, clinical, and physical characteristics, including age, biological sex, psychiatric diagnoses, and participants' body mass index (BMI), in a multiple linear regression.

To evaluate (2) between-person consistency in response scale interpretation, specifically what participants interpreted as neutral, before and after the EMA period, we categorized individuals' responses into three perceptions of neutrality: 1) *neutral*, indicating neutrality at the VAS midpoint (i.e., 45 to 55 for the unipolar item; -5 to 5 for the bipolar item), with a margin of 5 points above and below the respective midpoints (50 and 0); 2) *positive*, indicating neutrality using values above the VAS midpoint (i.e., 56 to 100 on unipolar and > 5 on bipolar), and 3) *negative*, indicating neutrality using values below the midpoint (i.e., 6 to 44 on unipolar and < -5 on bipolar). For the unipolar item, we included an additional category, 4) *zero*, for individuals who identified values between 0 and 5 as their neutral point. We then analyzed the overlap between participants' interpretation of neutrality on the two items by summarizing these categorizations in a cross-tabulation. We examined whether potential deviations from expected neutral values (45 to 55 on the unipolar and -5 to 5 on the bipolar VAS) could be correlated to the participant's person-specific mean on the respective item during the EMA-period.

To measure (3) test-retest consistency and response shifts across the two time points, before and after EMA, we calculated descriptive statistics for each situation and item (i.e., happy-self, happy-other, sad-self, sad-other) pre-EMA and post-EMA on affective items (happy and sad) the participants were repeatedly exposed to during the 30-day EMA period. Repeated exposure to items may lead to a response shift (e.g., changes in perception or desensitization to extreme end points during the EMA period), in which the post-EMA responses to the scenarios may differ from the pre-EMA responses. To investigate response shifts, we subtracted participants' pre-EMA ratings from each scenario from the corresponding post EMA-rating. We characterized the distribution of the change scores to determine the extent of response shifts across participants. The mean and standard deviation (SD) provide an estimate of the average shift and its variability across participants. The median indicates the tendency of this error; a median close to zero suggests that shifts are symmetrically distributed, without trends in shifting scores up or downwards after the EMA procedure. The interquartile range (IQR; 25th to 75th percentiles) reflects the spread of the middle 50% of the errors and highlights potential asymmetry or skewness in the distribution. Finally, the 95% percentiles (2.5th and 97.5th) offer insights into the distribution of extreme values, providing a more complete picture of response shifts, including potential outliers. We investigated the relationship between the individual change scores and demographic and clinical characteristics, including age, biological sex, psychiatric diagnoses, education levels, and employment status, in a multiple linear regression, with a Bonferroni-corrected α level set to .003 to adjust for multiple testing ($\alpha_{adjusted} = \frac{0.05}{16}$).

All analyses were conducted in R version 4.4.1 (R Core Team, 2024a). We used the *ggplot2* package (Wickham, 2016) for data visualization and the *stats* package (R Core Team, 2024b) for the regression analyses.

Code and Materials

Due to restrictions, we are precluded from sharing our data publicly. We provide all code and more information on the computational environment in the online supplement (osf.io/utdjq/).

Results

Participants were aged between 18 and 87 years (M = 47.09; SD = 14.41), with 2,511 of 3,761 (67.03%) being female.

Tactile Precision

On average, participants selected 31.01 (SD = 3.28, *range* = [11, 93]) when asked to select 31 on the 0-100 VAS. The majority of responses (87.0%) were between 30 and 32, and 59.0% were exactly 31. Figure 1 displays the distribution of responses overall, as well as by gender.

Among the demographic predictors, age was the only significant factor influencing tactile precision (*unstandardized* β = 0.018, t = 4.071, p < .001). No significant associations were found between tactile precision and sex (p = .797), psychiatric diagnosis (p = .807), or BMI (p = .573). Overall, only a small proportion of the variance (R^2 = 0.01,

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

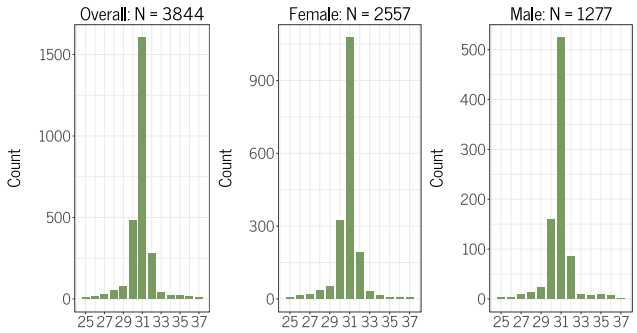


Figure 1. Distribution of Tactile Precision Stratified by Sex

Table 1. Cross-Table of Participants' Neutral Points Across Unipolar and Bipolar Item

		Unipolar				
		Zero (0 to 5)	*Negative* (6 to 44)	*Neutral* (45 to 55)	*Positive* (56 to 100)	**Total**
Bipolar	*Negative* < -5	0.11%	0.18%	0.07%	0.11%	**0.48%**
	Neutral -5 to 5	2.83%	4.53%	78.81%	2.72%	**88.89%**
	Positive > 5	0.15%	0.52%	3.31%	6.66%	**10.63%**
	Total	**3.09%**	**5.22%**	**82.19%**	**9.49%**	**100.00%**

Note. The margins contain the categorization of participants' perception of neutral points on the unipolar scale (bottom row) and the bipolar scale (rightmost column). The gray shaded row (bipolar) and column (unipolar) highlight the categorization of neutral on the two items.

$p < .001$) of the outcome was accounted for in the multiple regression.

Thus, we conclude that the effects for tactile imprecision in VAS ratings are negligible and that between-person differences in tactile precision are minimally explained by demographic and physical characteristics.

Between-Person Consistency in Perceptions of the Neutral Point

On both the unipolar and the bipolar scales, the large majority of individuals chose the midpoint of the scale as their neutral rating. For the unipolar (0 to 100) scale, 82.19% of participants identified their neutral point to be around the midpoint (45 to 55), with 68.87% indicating their neutral point to be exactly 50. For the bipolar (-50 to 50) scale, 88.89% of participants indicated the midpoint range (-5 to 5) as neutral point, and 78.84% identified exactly 0 as neutral point. Table 1 shows the categorization of people into the four levels of the unipolar scale, and the three levels of the bipolar scale. Among participants, 78.81% who selected values between -5 and 5 as neutral on the bipolar scale also chose values between 45 and 55 on the unipolar scale.

Figure 2 illustrates the distributions of participants' perceptions of neutral points across these two items. For the unipolar item (right), the majority of responses were around the midpoint and the distribution shows a roughly

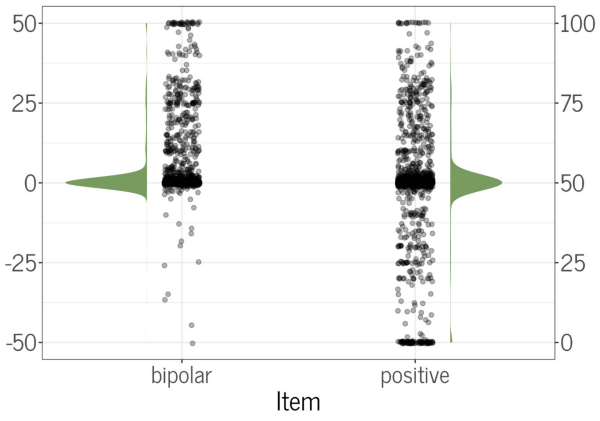


Figure 2. Distributions of Neutral Scale Ratings for Both Scales

equal spread of neutral values both above and below the midpoint. This suggests that participants' adjustments were similarly distributed in response to both positive and less positive interpretations of the neutral point. For the bipolar item (left), the majority of responses were around the midpoint (0), with distributions further showing a notable asymmetry (i.e., the tail extends more toward the positive end than the negative end). This suggests that participants were more likely to interpret neutral values in the positive range on the bipolar scale.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

To further understand the responses of the participants that fell outside the expected neutral ranges (45 to 55 on the unipolar and -5 to 5 on the bipolar scale), we conducted a post-hoc analysis. We investigated whether the interpretation of the neutral point of the 478 (18%) individuals who reported a neutral point that was lower than 45 or higher than 55 on the unipolar scale, and of the 295 (11%) of individuals who reported a neutral point on the bipolar scale outside the range of -5 to 5 was related to the person-specific mean of the respective item across the EMA period. We computed the within-person mean scores of the unipolar and bipolar items during the EMA period (i.e., mean scores of up to 120 observations per person) and correlated them with the reported neutral points in the post-EMA questionnaire for those respective items.

For the unipolar item, the correlation between the within-person means and the reported neutral points at post-EMA was r = .37 for the sample as a whole, but when looking specifically at individuals who reported an unexpected neutral value (< 45 or > 55), this correlation rose to r = .71. Similarly, for the bipolar item, the correlation between within-person EMA means and reported neutral points was r = .31 for the sample as a whole, and r = .73 for the subsample that answered outside the expected neutral range (< -5 and > 5). This suggests that the individuals who did not report within the expected neutral ranges may have related their answer more strongly to their personal average affect during the EMA period.

Test-Retest Consistency

Overall, the 16 affect scenario ratings showed a high average test-retest consistency from before to after the 30-day EMA period. The mean change was lowest for "On a very good vacation" (*Self:* M_{sad} = -0.37, SD_{sad} = 11.57; M_{happy} = -0.19, SD_{happy} = 13.74; *Other*: M_{sad} = -1.56, SD_{sad} = 14.97; M_{happy} = 0.75, SD_{happy} = 16.59) and highest for "Loosing Wallet" (*Self:* M_{sad} = -4.76, SD_{sad} = 20.69; M_{happy} = -1.83, SD_{happy} = 10.41; *Other*: M_{sad} = 6.54, SD_{sad} = 23.11; M_{happy} = -2.15, SD_{happy} = 11.60). Generally, change scores were larger for ratings of how others would feel compared to self-ratings.

For 12 out of 16 situation-affect ratings had a median change score of zero. For the remaining four items, the median change was between 2 to 5 points on a 0-100 scale (Table 2), which were all items inquiring about sadness, querying about how sad participants themselves would be or how sad they believed most others would be after a negative scenario.

The percentile ranges indicated notable individual variability in change scores. Specifically, the 50% interquartile range (IQR) was wider for sadness ratings in response to negative scenarios and for happiness ratings in positive scenarios. The widest spread was observed in the "Losing Wallet" vignette, where the IQR spanned from -7 to 17 (self-sad) and -6 to 19 (other-sad). In this vignette, the 95% range showed that for sadness and happiness ratings changed between -37 and 48 or even -37 and 56 points, indicating that in rare cases, participants' ratings shifted half the scale.

Together, these findings suggest that while median test-retest stability was strong (with little systematic bias), individual-level variability was larger for certain scenarios than others, particularly those involving negative events and sadness ratings or incongruent pairings. Among the demographic variables, only age was a significant predictor of change in responses from pre- to post-EMA assessment on two of the 16 variables, with higher age related to less change on "feeling happy despite losing one's wallet" (*unstandardized* β = -0.07, p < .001, R^2 = .007) and perception of "others feeling happy when losing their wallet" (*unstandardized* β = -0.09, p < .001, R^2 = .011). Using the largest effect sizes (-0.09) as an example, this indicates that being 20 years older predicts changing 1.8 points less from pre- to post-EMA assessment.

We did not pre-specify inference effect sizes or equivalence bounds; therefore, we cannot formally conclude whether the observed changes constitute meaningful response shifts. However, in a recent test-retest reliability study, changes of up to 8 points were attributed to measurement error alone, suggesting that the observed shifts in our study fall well below this threshold (Dejonckheere et al., 2022).

Discussion

VAS has become mainstream in EMA and the electronic administration of questionnaires, given the ease of implementation in a graphical interface and desire to measure variables in more continuous intervals. However, there is minimal psychometric examination of VAS ratings. In this randomly sampled nationwide study of Norwegian adults, we demonstrated, using three complementary methods, that participants accurately and consistently use VAS scales, with findings that can be interpreted within the scope and limitations of these methods. First, we found remarkably high tactile precision across participants in their ability to select specific numbers on VAS. Second, we found high consistency across participants in what they perceived as a "neutral" point of a VAS slider on two different types of VAS scales (unipolar and bipolar), although there was some variation. Third, we found high test-retest consistency within participants over time in their anticipated affective response to positive and negative affective scenarios. Each of these findings warrants further discussion.

In our test of (1) tactile precision, participants were highly accurate in selecting the specified number ('31'), and this did not meaningfully differ across sex, presence of psychiatric diagnoses, BMI, or age. These results alleviate concerns about a specific and practical domain of measurement error resulting from physical imprecision. The accuracy of 87% (scoring 31 ± 2 points) was even higher than that found in paper-based VAS scales, where only 50% of responses were accurate up to ± 2 points (i.e., 2 mm) and 90% of responses were accurate up to ± 9 points (Bijur et al., 2001). Older people were slightly less precise in their use of the scale, although the effect size was very small.

We also examined the (2) interpretation of the response scale across participants and whether they used the same number as neutral score. Overall, a clear majority of partic-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

Table 2. Pre-Post Changes in Situation Ratings

Situation / Emotion		Mean (Standard Deviation)			Median			Inter Quartile Range			95%		
		Pre	Post	Change	Pre	Post	Change	Pre	Post	Change	Pre	Post	Change
Serious Argument													
Self	Sad	77.53 (18.20)	80.41 (17.75)	2.66 (17.75)	80	84	2	70, 90	73, 93	-6, 11	29, 100	29.92, 100	-33, 40
	Happy	4.99 (11.36)	3.28 (9.36)	-1.73 (12.10)	0	0	0	0, 5	0, 1	-3, 0	0, 33	0, 26	-23, 16
Other	Sad	72.45 (17.25)	76.62 (15.72)	3.65 (18.40)	75	79	3	64, 84	69, 87	-6, 13	29, 100	39.92, 100	-33, 42.17
	Happy	7.68 (13.59)	5.07 (11.02)	-2.69 (14.44)	1	0	0	0, 10	0, 6	-5, 0	0, 51	0, 36	-37, 23.17
Lost Wallet													
Self	Sad	52.99 (25.48)	58.25 (25.47)	4.76 (20.96)	55	62	4	31, 71	43, 76	-7, 17	5, 100	1.85, 100	-37.17, 48
	Happy	4.03 (9.48)	2.37 (6.98)	-1.83 (10.41)	0	0	0	0, 3	0, 0	-2, 0	0, 32	0, 25	-27.17, 15
Other	Sad	57.80 (23.36)	65.07 (21.58)	6.54 (23.11)	60	70	5	46, 75	53, 80	-6, 19	4, 100	12, 100	-37, 56.17
	Happy	4.89 (10.91)	2.81 (7.90)	-2.15 (11.60)	0	0	0	0, 5	0, 0	-3, 0	0, 40	0, 27	-31, 18
Very Good Vacation													
Self	Sad	4.88 (11.17)	4.52 (10.99)	-0.37 (11.57)	0	0	0	0, 5	0, 3	-1, 0	0, 36	0, 37.07	-21.17, 21
	Happy	86.03 (14.70)	85.94 (14.64)	-0.19 (13.74)	89	89	0	80, 100	79, 100	-6, 6	50, 100	50, 100	-25, 26
Other	Sad	4.79 (13.07)	3.21 (10.16)	-1.56 (14.97)	0	0	0	0, 4	0, 1	-2, 0	0, 47	0, 24	-27, 17
	Happy	86.02 (15.23)	86.93 (13.19)	0.75 (16.49)	90	89	0	80, 98	80, 99	-6, 7	50, 100	57, 100	-26, 33.17
Time with Friends													
Self	Sad	6.22 (11.96)	4.95 (11.44)	-1.23 (11.43)	0	0	0	0, 8	0, 4	-3, 0	0, 42	0, 40	-22, 21
	Happy	81.30 (14.99)	83.54 (14.55)	1.75 (13.25)	82	85	0	73, 92	76, 95	-5, 9	48, 100	50, 100	-23, 27
Other	Sad	4.69 (10.76)	3.15 (8.87)	-1.47 (11.86)	0	0	0	0, 6	0, 2	-3, 0	0, 27	0, 22	-19, 14
	Happy	83.29 (14.05)	85.49 (12.19)	1.76 (14.54)	85	87	0	77, 93	79, 95	-5, 8	51, 100	59.92, 100	-23, 29

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

ipants demonstrated strong consistency when asked to select the 'neutral' response on two commonly used types of VAS scales (unipolar and bipolar). Specifically, 82% of participants indicated the neutral point to be in the range of 45 to 55 for the unipolar (0 to 100) VAS, and 89% indicated the neutral point to be between -5 and 5 for the bipolar (-50 to 50) VAS. However, this also implies that 18% (for the unipolar item) and 11% (for the bipolar item) selected a 'neutral' response outside of the expected range, suggesting some variability in how participants interpreted the response scale. We also found that there was a small number (< 3%) of people who indicated 0 as their neutral point on the unipolar scale. In our post hoc analyses, we found that values provided by individuals who responded outside the expected neutral ranges were strongly correlated with their average rating during the EMA period. However, more research is needed to identify other potential individual characteristics impacting scale interpretation. This suggests that a small proportion of adults may have interpreted neutral values on VAS scales to correspond to how they usually tended to feel.

In our sample, 22% of participants differed in the identification of the neutral point on at least one of the two scales, indicating variation and ambiguity in VAS interpretation. Unlike Likert scales, which typically provide an explicit neutral category, VAS formats do not offer intermediate labels or anchors. However, the use of a neutral point in Likert scales has also been debated. DeCastellarnau (2018) summarized the evidence, highlighting that providing a neutral option can encourage satisficing and avoidance of attitude expression (Kulas & Stachowski, 2009) but can also reduce extreme response styles (Weijters et al., 2010). Conversely, removing a neutral option (e.g., in even-numbered Likert scales) has been shown to increase forced-choice responding and participant frustration (Rozin & Royzman, 2001). Including a neutral point may help participants anchor their judgments relative to a "typical" feeling, which may have happened here as we found relations with the typical level of affect across the EMA scales (Tourangeau et al., 2004). Relatedly, in recent research on EMA scales, Dejonckheere et al. (2023) found that providing a relative anchor (e.g., the last rating) made it easier for participants to use the scale and led to greater perceived accuracy.

In the current evaluation, we focused specifically on the interpretation of the neutral point. However, ambiguity likely extends beyond neutrality: especially in unipolar VAS scales, the interpretation of the full continuum of intensity (from low to high) may also vary across individuals. Future research should investigate the extent to which participants' perceptions of neutral values differ on Likert scales, where neutral values sometimes are explicitly defined, compared to VAS scales, where these values are often not defined.

Finally, we (3) tested whether participants demonstrated shifts in their interpretation of the response scale for a positive (happy) and negative (sad) affect item after a 30-day EMA period in which participants repeatedly rated their momentary happiness and sadness. We found high levels of test-retest consistency across all items. While the mean and median change scores were close to 0 in the overall sample level, the spread around these values highlights that there are subgroups of individuals who demonstrate greater response shifts. This provides an important avenue for future research to identify which individuals experience such shifts and better understand what these shifts are related to. Notably, previous work demonstrated response shifts are often identified post-hoc using study designs where the construct of interest is *expected* to change over time (measurement of negative affect in a treatment sample; Kramer et al., 2014). Our study was designed to examine the response shift in constructs of interest that were *not* expected to change over time; thus, our results provide a useful test of response shift and add to the relatively small literature that directly tests response shift (see Eisele et al., 2023, for a prior example). Overall, these findings reflect a helpful first step towards developing a more thorough understanding of the influence of repeated assessment on response shift.

Strengths and Limitations

This study has several strengths. First, the use of a nationwide random sample contributes to a broad representation of individuals across different subgroups of adults, enhancing the generalizability of the findings to the general adult population. Second, the large sample size, reflecting one of the largest EMA studies conducted to date, contributes to a robust evaluation of key psychometric concerns raised in the literature. Third, we investigated between-person consistency in scale interpretation using two types of commonly employed VAS scales (unipolar and bipolar) in the literature. Measuring affect with single items is favored by many researchers for its ease and validity (Cloos et al., 2023; Dejonckheere et al., 2022); therefore, information on such measures is highly relevant to the field. Moreover, test-retest consistency and response shifts were evaluated after extensive exposure to affective items (up to 120 times per person) during the EMA period and through pre- and post-EMA test-retest comparisons on a large set of items (16), allowing for a detailed assessment of potential changes in participants' responses over time and across different types of affective categories (both positive and negative affect). The employment of a balanced set of emotion items is important to discern differences in response shifts (Arslan et al., 2020).

This study also includes several limitations. The VAS ratings investigated in this study were conducted in the context of EMA using smartphones, which introduces several limitations. First, tactile precision may have been influenced by device type and screen size, as small screens or other physical constraints can affect participants' ability to make accurate selections on a VAS slider (Van Berkel et al., 2020). The findings warrant replication with a wider array of numbers, particularly near the end of the scale (e.g., 10 and 90), as the use of larger screens can impair the precision at the edge of the screen because it is harder to reach further to the edge (of the left or right, depending on the person) with the thumb (Karlson et al., 2006). Additional

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

information on the handedness of participants may provide information on these differences in future studies.

Second, the design of our tactile item may have cued participants to put a greater effort into tapping a requested number. It did not account for the preceding cognitive response processes required to decide which number to tap (Hubley, 2021) or the interaction between deciding and tapping. While the physical process by which participants can tap a pre-defined number is a necessary condition for accurate measurement, it is not sufficient. Additionally, tactile precision was assessed following the EMA period and thus was not administered to participants who withdrew from the study early. Thus, our assessment of tactile precision may be biased towards individuals who were more likely to respond with higher precision, and additional tests of tactile precision, especially earlier on in the EMA period, are warranted.

Third, the between-person differences in interpreting the neutral point, as related to the mean EMA score, reflect an interpretation based on the data. However, it remains unclear whether participants genuinely interpreted the neutral as their average or baseline or if they misunderstood the assignment. Factors such as motivation and attention levels may have also influenced response accuracy, as participants might have been less focused or engaged. Moreover, the test-retest consistency design only allows for investigations of differences between pre- and post-EMA and cannot tell us about when the shift happened. Finally, our analysis did not separate measurement error, precluding insights into to what extent changes in responses were related to measurement error rather than conceptual changes.

Lastly, this study focused exclusively on VAS scales and did not include a direct comparison with Likert-type scales. As a result, the findings are only interpretable within the context of VAS methodology and should not be taken to imply that VAS is generally superior to Likert scales. In the survey literature, this comparison has been explored more extensively; for example, DeCastellarnau (2018) provides a comprehensive review of differences between VAS and Likert scales in the context of survey measures.

Future Research

Moving forward, both direct replications (e.g., with different participants or in other languages) and conceptual replications (e.g., exploring response scale interpretations of anchor points or assessing other affective or neutral imagined scenarios) are needed to gain insights into the accuracy and the validity of VAS in digital assessments.

For instance, we know relatively little about the internal processes participants engage in before selecting a numerical value, such as how they interpret a scale, reflect on their current state (e.g., happiness), and translate that experience into a specific number. These internal response processes are central to the validity of ESM data but are rarely examined directly.

Future research should therefore prioritize efforts to uncover these mechanisms. Mixed-methods approaches, such as cognitive interviewing, think-aloud protocols, or real-time verbal reports, are especially well-suited to gain insight into the specific cognitive processes underlying how individuals select their responses (Truijens et al., 2023). Similarly, such methods could help determine whether individuals understand scale anchors (e.g., the neutral midpoint or the extremes) in a consistent way and whether these interpretations vary across participants or change over time.

Furthermore, response processes, what people do, think, or feel when interacting with the item (Hubley & Zumbo, 2017), can shift over time, with implications for the validity of EMA data. Investigating such shifts, both through statistical modeling and qualitative inquiry, is essential to determine the nature and extent of potential measurement artifacts in EMA data (Eisele et al., 2025; Schreuder et al., 2020). There are methods available to examine whether the measurement model remains invariant over time (Vogelsmeier et al., 2023) and whether participants' understanding of the latent construct shifts during data collection (Vogelsmeier et al., 2024).

Finally, while the question of whether VAS or Likert scales are more appropriate has received growing attention in recent years, further research, including direct comparison, is needed to determine which format is better suited for use in EMA settings. Drawing such conclusions requires careful consideration of how each scale type functions in capturing momentary experiences in naturalistic contexts.

Conclusion

Taken together, our results suggest that participants are generally accurate and consistent when selecting specific numbers on VAS items and choosing neutral response options, and we found no evidence for response shifts when responding to affective items before and after repeated prompts about their emotions in daily life. These findings alleviate key concerns regarding physical inaccuracy and inconsistencies both between and within individuals in using VAS scales for digital self-reporting. Ultimately, this study provides valuable insights for researchers designing digital assessment studies (e.g., *item* formulation versus *response option* formulation) and highlights areas for future psychometric work to advance digital assessment tools.

Funding

LC was funded by grant PDMT2/24/034 from the Research Council of KU Leuven.

MLP received funding from the National Institute on Alcohol Abuse and Alcoholism (R00AA029459).

Acknowledgements

This paper was conceptualized at the workshop "It's about Time - Improving Intensive Longitudinal Data" organized by the "Measurement is the New Black"-consortium at the Lorentz Center in Leiden, the Netherlands in November 2024.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

Data Availability

The data is currently not publicly available, we have made a synthetic dataset based on our exact dataset with the same properties, which allow replication of our analysis, inspection of our results and data. The synthetic data and code are available in the computational environment online (osf.io/utdjq/).

Conflict of Interest

The authors declare that they have no competing interests.

Author contributions

LC: Conceptualization; Formal analysis; Visualization; Writing – original draft; Writing – review and editing

BSS: Conceptualization; Formal analysis; Visualization; Writing – original draft; Writing – review and editing

MLP: Conceptualization; Formal analysis, Writing - original draft; Writing - review and editing

EF: Conceptualization; Formal analysis, Writing - original draft; Writing - review and editing

SBW: Conceptualization; Formal analysis, Writing - original draft; Writing - review and editing

MAH: Conceptualization; Data curation; Formal analysis; Methodology; Investigation; Writing – review and editing

SUJ: Investigation; Writing – review and editing

AH: Investigation; Writing – review and editing

OVE: Conceptualization; Methodology; Formal analysis; Investigation; Data curation; Project administration; Writing – original draft; Writing – review and editing; Supervision

Editors: Andrea Howard (Senior Editor)

Submitted: February 13, 2025 PDT. Accepted: June 27, 2025 PDT. Published: August 06, 2025 PDT.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at http://creativecommons.org/licenses/by/4.0 and legal code at http://creativecommons.org/licenses/by/4.0/legalcode for more information.

References

Abend, R., Dan, O., Maoz, K., Raz, S., & Bar-Haim, Y. (2014). Reliability, validity and sensitivity of a computerized visual analog scale measuring state anxiety. *Journal of Behavior Therapy and Experimental Psychiatry*, *45*(4), 447–453. https://doi.org/10.1016/j.jbtep.2014.06.004

Aitken, R. C. B. (1969). Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine*, *62*(10), 989–993. https://doi.org/10.1177/003591576906201005

Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000699

Anvari, F., Efendic, E., Olsen, J., Arslan, R. C., Elson, M., & Schneider, I. K. (2023). Bias in self-reports: An initial elevation phenomenon. *Social Psychological and Personality Science*, *14*(6), 727–737. https://doi.org/10.1177/19485506221129160

Arslan, R. C., Reitz, A. K., Driebe, J. C., Gerlach, T. M., & Penke, L. (2020). Routinely randomize potential sources of measurement reactivity to estimate and adjust for biases in subjective reports. *Psychological Methods*. https://doi.org/10.1037/met0000294

Bijur, P. E., Silver, W., & Gallagher, E. J. (2001). Reliability of the visual analog scale for measurement of acute pain. *Academic Emergency Medicine*, *8*(12), 1153–1157. https://doi.org/10.1111/j.1553-2712.2001.tb01132.x

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

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Cloos, L., Ceulemans, E., & Kuppens, P. (2023). Development, validation, and comparison of self-report measures for positive and negative affect in intensive longitudinal research. *Psychological Assessment*, *35*(3), 189. https://doi.org/10.1037/pas0001200

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity*, *52*(4), 1523–1559. https://doi.org/10.1007/s11135-017-0533-4

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. https://doi.org/10.1037/pas0001178

Dejonckheere, E., Penne, I., Briels, L., & Mestdagh, M. (2023). For better or for worse? Visualizing previous intensity levels improves emotion (dynamic) measurement in experience sampling. *Psychological Assessment*. https://doi.org/10.1037/pas0001296

Eisele, G., Hiekkaranta, A., Kunkels, Y. K., Rot, M. aan het, van Ballegooijen, W., Bartels, S. L., Bastiaansen, J. A., Beymer, P. N., Bylsma, L. M., Carpenter, R. W., Ellison, W. D., Fisher, A. J., Forkmann, T., Frumkin, M. R., Fulford, D., Naragon-Gainey, K., Greene, T., Heininga, V. E., Jones, A., … Kirtley, O. J. (2025). ESM-Q: A consensus-based quality assessment tool for experience sampling method items. *Behavior Research Methods*, *57*(4), 124. https://doi.org/10.3758/s13428-025-02626-1

Eisele, G., Vachon, H., Lafit, G., Tuyaerts, D., Houben, M., Kuppens, P., Myin-Germeys, I., & Viechtbauer, W. (2023). A mixed-method investigation into measurement reactivity to the experience sampling method: The role of sampling protocol and individual characteristics. *Psychological Assessment*, *35*(1), 68–81. https://doi.org/10.1037/pas0001177

Fritz, J., Piccirillo, M., Cohen, Z. D., Frumkin, M., Kirtley, O., Moeller, J., Neubauer, A., Norris, L., Schuurman, N. K., Snippe, E., & Bringmann, L. (2023). *So you want to do ESM? Ten essential topics for implementing the Experience Sampling Method (ESM)*. OSF. https://doi.org/10.31219/osf.io/fverx

Funke, F., & Reips, U.-D. (2012). Why Semantic Differentials in Web-Based Research Should Be Made from Visual Analogue Scales and Not from 5-Point Scales. *Field Methods*, *24*(3), 310–327. https://doi.org/10.1177/1525822X12444061

García-Pérez, M. A., & Alcalá-Quintana, R. (2023). Accuracy and precision of responses to visual analog scales: Inter- and intra-individual variability. *Behavior Research Methods*, *55*(8), 4369–4381. https://doi.org/10.3758/s13428-022-02021-0

Gunthert, K. C., & Wenze, S. J. (2012). Daily diary methods. In *Handbook of research methods for studying daily life* (pp. 144–159). The Guilford Press.

Haslbeck, J., Martínez, A. J., Roefs, A., Fried, E. I., Lemmens, L. H. J. M., Groot, E., & Edelsbrunner, P. (2025). *Comparing Likert and Visual Analogue Scales in Ecological Momentary Assessment*. OSF. https://doi.org/10.31234/osf.io/yt8xw_v2

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

Hubley, A. M. (2021). Response processes validity evidence: Understanding the meaning of scores from psychological measures. In *Handbook on the state of the art in applied psychology* (pp. 413–434). Wiley Blackwell.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

Hubley, A. M., & Zumbo, B. D. (2017). Response Processes in the Context of Validity: Setting the Stage. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 1–12). Springer International Publishing. https://doi.org/10.1007/978-3-319-56129-5_1

Junghaenel, D. U., & Stone, A. A. (2020). Ecological momentary assessment for the psychosocial study of health. In *The Wiley Encyclopedia of Health Psychology* (pp. 105–112). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119057840.ch56

Karlson, A. K., Bederson, B., & Contreras-Vidal, J. (2006). *Understanding single-handed mobile device interaction*. https://www.semanticscholar.org/paper/Understanding-Single-Handed-Mobile-Device-Karlson-Bederson/67702fe26fabeaf28df44837ea28d153fce3c1b3

König, L. M., Allmeta, A., Christlein, N., Van Emmenis, M., & Sutton, S. (2022). A systematic review and meta-analysis of studies of reactivity to digital in-the-moment measurement of health behaviour. *Health Psychology Review*, *16*(4), 551–575. https://doi.org/10.1080/17437199.2022.2047096

Kramer, I., Simons, C. J. P., Hartmann, J. A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., Schruers, K., van Bemmel, A. L., Myin-Germeys, I., Delespaul, P., van Os, J., & Wichers, M. (2014). A therapeutic application of the experience sampling method in the treatment of depression: A randomized controlled trial. *World Psychiatry*, *13*(1), 68–77. https://doi.org/10.1002/wps.20090

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, *43*(3), 489–493. https://doi.org/10.1016/j.jrp.2008.12.005

Mestdagh, M., & Dejonckheere, E. (2021). Ambulatory assessment in psychopathology research: Current achievements and future ambitions. *Current Opinion in Psychology*, *41*, 1–8. https://doi.org/10.1016/j.copsyc.2021.01.004

Myin-Germeys, I., & Kuppens, P. (Eds.). (2022). *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (2nd ed.). Center for Research on Experience Sampling and Ambulatory Methods. https://www.kuleuven.be/samenwerking/real/real-book/index.htm

R Core Team. (2024a). *R: a language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

R Core Team. (2024b). *R: the R stats package* [Computer software]. R Foundation for Statistical Computing. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html

Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, *40*(3), 699–704. https://doi.org/10.3758/BRM.40.3.699

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Saltzman, L. Y., Terzis, L. D., Hansel, T. C., Blakey, J. M., Logan, D., & Bordnick, P. S. (2021). Harnessing technology for research during the COVID-19 pandemic: A mixed methods diary study protocol. *International Journal of Qualitative Methods*, *20*, 1609406920986043. https://doi.org/10.1177/1609406920986043

Schreuder, M. J., Groen, R. N., Wigman, J. T. W., Hartman, C. A., & Wichers, M. (2020). Measuring psychopathology as it unfolds in daily life: Addressing key assumptions of intensive longitudinal methods in the TRAILS TRANS-ID study. *BMC Psychiatry*, *20*(1), 351. https://doi.org/10.1186/s12888-020-02674-1

Schwartz, C. E., Sprangers, M. A. G., Carey, A., & Reed, G. (2004). Exploring response shift in longitudinal data. *Psychology & Health*, *19*(1), 51–69. https://doi.org/10.1080/0887044031000118456

Setnik, B., Roland, C. L., Pixton, G., & Webster, L. (2017). Measurement of Drug Liking in Abuse Potential Studies: A Comparison of Unipolar and Bipolar Visual Analog Scales. *The Journal of Clinical Pharmacology*, *57*(2), 266–274. https://doi.org/10.1002/jcph.801

Shrout, P. E., Stadler, G., Lane, S. P., Joy McClure, M., 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 of the United States of America*, *115*(1), E15–E23. https://doi.org/10.1073/pnas.1712277115

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557–566. https://doi.org/10.1037/pas0000648

Stinson, L., Liu, Y., & Dallery, J. (2022). Ecological momentary assessment: A systematic review of validity research. *Perspectives on Behavior Science*, *45*. https://doi.org/10.1007/s40614-022-00339-w

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

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, *68*(3), 368–393. https://doi.org/10.1093/poq/nfh035

Truijens, F. L., De Smet, M. M., Vandevoorde, M., Desmet, M., & Meganck, R. (2023). What is it like to be the object of research? On meaning making in self-report measurement and validity of data in psychotherapy research. *Methods in Psychology*, *8*, 100118. https://doi.org/10.1016/j.metip.2023.100118

van Berkel, N. (2017). The experience sampling method on mobile devices. *ACM Comput. Surv*, *50*(6), 1–40. https://doi.org/10.1145/3123988

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

Van Berkel, N., Goncalves, J., Wac, K., Hosio, S., & Cox, A. L. (2020). Human accuracy in mobile data collection. *International Journal of Human-Computer Studies*, *137*, 102396. https://doi.org/10.1016/j.ijhcs.2020.102396

Vogelsmeier, L. V. D. E., Cloos, L., Kuppens, P., & Ceulemans, E. (2023). Evaluating dynamics in affect structure with latent Markov factor analysis. *Emotion*. https://doi.org/10.1037/emo0001307

Vogelsmeier, L. V. D. E., Jongerling, J., & Maassen, E. (2024). Assessing and accounting for measurement in intensive longitudinal studies: Current practices, considerations, and avenues for improvement. *Quality of Life Research*, *33*(8), 2107–2118. https://doi.org/10.1007/s11136-024-03678-0

Warriner, A. B., Shore, D. I., Schmidt, L. A., Imbault, C. L., & Kuperman, V. (2017). Sliding into happiness: A new tool for measuring affective responses to words. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Experimentale*, *71*(1), 71–88. https://doi.org/10.1037/cep0000112

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Wenz, A., & Keusch, F. (2023). Increasing the acceptance of smartphone-based data collection. *Public Opinion Quarterly*, *87*(2), 357–388. https://doi.org/10.1093/poq/nfad019

Wickham, H. (2016). *Ggplot2*. Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Yeung, A. W. K., & Wong, N. S. M. (2019). The Historical Roots of Visual Analog Scale in Psychology as Revealed by Reference Publication Year Spectroscopy. *Frontiers in Human Neuroscience*, *13*. https://doi.org/10.3389/fnhum.2019.00086

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025

Supplementary Materials

Peer Review Communication

Download: https://collabra.scholasticahq.com/article/142735-accuracy-and-consistency-of-visual-analog-scales-in-ecological-momentary-assessment-and-digital-studies/attachment/296058.docx?auth_token=KjRtUV0jSih3O9Ss-nU4

Downloaded from http://online.ucpress.edu/collabra/article-pdf/11/1/142735/890934/collabra_2025_11_1_142735.pdf by guest on 18 December 2025