Original Research Article

Assessment
1–19
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10731911241283623
journals.sagepub.com/home/asm
Sage

Using Freely Generated Labels Instead of Rating Scales to Assess Emotion in Everyday Life

Katie Hoemann[1], Evan Warfel[2], Caitlin Mills[3], Laura Allen[3], Peter Kuppens[1], and Jolie B. Wormwood[4]

Abstract

To measure emotion in daily life, studies often prompt participants to repeatedly rate their feelings on a set of prespecified terms. This approach has yielded key findings in the psychological literature yet may not represent how people typically describe their experiences. We used an alternative approach, in which participants labeled their current emotion with at least one word of their choosing. In an initial study, estimates of label positivity recapitulated momentary valence ratings and were associated with self-reported mental health. The number of unique emotion words used over time was related to the balance and spread of emotions endorsed in an end-of-day rating task, but not to other measures of emotional functioning. A second study tested and replicated a subset of these findings. Considering the variety and richness of participant responses, a free-label approach appears to be a viable as well as compelling means of studying emotion in everyday life.

Keywords

emotions, language, experience sampling, mental health, individual differences, qualitative methods

Emotions take a lot of forms in everyday life: You can be "excited," "disappointed," and "bored"; you can also be "really looking forward to it," "kinda bummed," "meh," and much more. Studies seeking to capture these moments commonly ask people to report on their current experience over time—an approach known alternatively as experience sampling (Csikszentmihalyi & Larson, 1987) or ecological momentary assessment (Stone & Shiffman, 1994)—by rating the intensity of their feelings using Likert-type-style or visual analog (i.e., slider) scales. These ratings may be on affective dimensions such as how pleasant the participant feels (i.e., valence; Russell, 1980), or on a set of prespecified emotions (see Watson et al., 1988). Such structured, quantitative assessments can be used to model between-person differences and within-person fluctuations in emotion and have yielded key findings in the psychological literature (for review and discussion, see Kuppens et al., 2022). At the same time, rating scales may not represent how people typically think about or describe their feelings, and may miss important sources of nuance and variation (see Y. Li et al., 2020). Here, we report on an alternative approach in which participants were asked to describe their current experience with at least one word of their choosing. We illustrate how these self-generated emotion labels provide valuable, qualitative insight into people's daily lives. We then examine whether they also recapitulate momentary valence ratings and whether they are related to person-level measures of mental health and emotional functioning.

Using rating scales to measure emotional experience has a long and venerable history, starting with Nowlis' checklist (Nowlis, 1965; Nowlis & Nowlis, 1956) and the Profile of Mood States (POMS; McNair et al., 1971), and has formed the basis for studying emotion in everyday life (see Csikszentmihalyi & Figurski, 1982; Diener et al., 1984; Stone, 1981). One focus of this research has been modeling temporal trends within individuals and testing whether shifts in these dynamics are associated with fluctuations in stress, coping, and decision-making (Erbas et al., 2018, 2021; Tomko et al., 2015), as well as vulnerability toward and recurrence of mental ill-health (see Pe et al., 2015; Snippe et al., 2023; van de Leemput

[1]KU Leuven, Belgium
[2]University of California, Davis, USA
[3]University of Minnesota, Minneapolis, USA
[4]University of New Hampshire, Durham, USA

Corresponding Author:
Katie Hoemann, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, Box 3727, Leuven 3000, Belgium.
Email: khoemann@gmail.com

et al., 2014). A second focus has been quantifying individual differences in affective and emotional experiences as they manifest in context. Studies have shown, for example, that people differ in their mean or tonic valence (see R. J. Larsen & Diener, 1987) and in how much their affect tends to vary (see R. J. Larsen, 1987; Mestdagh et al., 2018) or linger over time (see inertia; Kuppens et al., 2010). Ratings of specific emotions make it possible to study their distributions and inter-relationships. Two relevant individual differences are the range and balance of emotions experienced—known as emodiversity (Quoidbach et al., 2014)—and the context-specificity with which these experiences occur—known as emotional granularity or emotion differentiation (Barrett et al., 2001; Tugade et al., 2004). Higher emodiversity and emotional granularity have been widely associated with wellness-related outcomes, including fewer mental health symptoms (see Demiralp et al., 2012; Quoidbach et al., 2014; for discussion, see Dejonckheere et al., 2019) and more adaptive coping (see Grossmann et al., 2019; Pond et al., 2012; for reviews, see O'Toole et al., 2020; Seah & Coifman, 2021). In sum, repeated momentary ratings of emotion terms have provided critical insights into within- and between-person variation in affective and emotional experiences.

While experience sampling methods enhance ecological validity by capturing emotional experiences and trends inaccessible in standardized laboratory settings (Wilhelm & Grossman, 2010), the use of closed-ended word lists and rating scales inherently limits the variation that can be observed. Structured emotion ratings may not capture the heterogeneity and nuance of everyday feelings equally for all people. Standard lists of emotion terms are likely sufficient for some, but people can also have experiences that are either not among or are poorly described by the emotion words provided (e.g., "bored" is not included in the Positive and Negative Affect Schedule [PANAS]; Y. Li et al., 2020)—and this may be especially true for individuals experiencing more diverse, precise, or complex emotions. The breadth and specificity of emotions covered can be increased by using more comprehensive lists of words, but this comes with great costs to participant burden and data quality (Eisele et al., 2020). More generally, the presence of specific words may shape participants' experience via order and priming effects, making certain emotions more accessible or salient than they otherwise would have been (see Hansen & Shantz, 1995). This can benefit people by increasing their awareness of different types of emotions (see Hoemann, Barrett, & Quigley, 2021; Widdershoven et al., 2019) but also carries potential impacts for measurement reactivity (Eisele et al., 2022). Collecting ratings of valence or other affective dimensions instead of a list of specific emotions is one way of addressing these considerations but necessarily reduces the richness of information gathered about emotional experience.

For these reasons, there is growing interest in complementing structured ratings with open-ended methods that impose fewer constraints on how people interpret and report on their emotions (Kashdan et al., 2015). Several of these methods involve analyzing the emotion words people use spontaneously in written or spoken descriptions of experience. Manual annotation schemes assessing the complexity of these descriptions (see Labouvie-Vief & Medler, 2002; Lane et al., 1990) have shown, for example, that people who report difficulty identifying and describing their emotions (i.e., who are higher in alexithymia) also use a less diverse set of emotion words (Wotschack & Klann-Delius, 2013) and that the use of more specific emotion words—understood as a measure of emotional granularity—is associated with greater daily life satisfaction (Ottenstein & Lischetzke, 2019). Recently, Vine and colleagues (2020) developed a means of automatically calculating the ratio of unique emotion words in expressive writing, finding that a larger variety of negative words was associated with worse well-being, and a larger variety of positive words was associated with better well-being. Automated approaches have also been used to estimate the valence of natural language descriptions by counting the proportion of positively and negatively valenced words (most commonly using the program Linguistic Inquiry and Word Count [LIWC]; Pennebaker et al., 2015) or applying other forms of sentiment analysis (for a review, see Mehta & Pandya, 2020). These language-derived valence estimates are associated, between persons, with measures of well-being such as symptoms of mental ill-health (see Tov et al., 2013; for a summary, see Sun et al., 2019, Table 1).

The above studies demonstrate that open-ended methods can be used to measure individual differences relevant to emotion and well-being. Research has also looked at whether the language people use reliably captures within-person fluctuations in emotional experience. For example, asking participants to produce the 10 words that best describe their recent emotions has been shown to yield a measure of positive and negative affects on par with established scales (see PANAS; Y. Li et al., 2020). Studies using natural language to model emotion dynamics have been more mixed in their results. Analyses of social media posts have shown a correspondence between language and momentary affect when the language features included are complex and the affect is manually annotated by human judges (see Eichstaedt & Weidman, 2020). In contrast, simple percentages of the positive and negative words used on Facebook (calculated using LIWC) do not consistently track the affect reported by participants themselves in asynchronous experience sampling (Kross et al., 2019). Analyses of ambient speech captured during experience sampling (i.e., what participants and those around them were saying) likewise found that LIWC-estimated valence was not related to contemporaneous participant ratings (Sun et al., 2019). Importantly, however, none of these studies of natural language used descriptions of current experience elicited from participants, leaving open the possibility that direct and intentional reports will index participant-reported affect from moment to moment. Recent work examining language use when participants were asked to describe what they had been doing and how it made them feel showed equivocal findings in this respect (Carlier et al., 2021). Yet examining this question with shorter-format descriptions (i.e., individual labels) may provide a more fruitful approach to validating language-based methods, and to potentially deploying them across a larger range of (clinical) contexts.

We continue the work on open-ended methods of capturing emotion in everyday life with two experience sampling studies in which participants used freely generated labels, rather than structured emotion ratings, to report on their momentary experience. In Study 1, we re-analyzed data from Hoemann et al. (2020), in which participants provided at least one word of their choosing and rated their valence at each prompt and completed measures of mental health and emotional functioning outside of experience sampling. We used these data to accomplish three goals. First, we took an inventory of the emotion labels used by participants, illustrating their nature and diversity across the sample as well as quantifying their number and uniqueness per prompt and per person. Second, we examined whether emotion labels recapitulated structured valence ratings both between and within persons. Third, we tested emotion labels' between-person convergent validity by examining whether measures of label valence, number, and uniqueness were related to mental health (following, see Vine et al., 2020) as well as to emotional functioning in terms of emodiversity (Quoidbach et al., 2014), alexithymia (Wotschack & Klann-Delius, 2013), emotional granularity (Ottenstein & Lischetzke, 2019), and emotional complexity (i.e., self-reported range and differentiation of emotional experience; Kang & Shaver, 2004). In Study 2, we sought to replicate a subset of our findings through a preregistered secondary analysis (https://osf.io/xceu5) of a separate experience sampling data set. This data set was collected as part of a larger study on affect and cognition in daily life that did not contain measures of emotional functioning, and so our tests of convergent validity focused only on emotion labels' relationship with mental health.

Study 1

Methods

Study 1 has been reported in detail by Hoemann et al. (2020; see also Hoemann, Barrett, & Quigley, 2021; Hoemann, Khan, et al., 2021; Hoemann, Lee, et al., 2023). Below we report aspects relevant for the present analyses, including all data exclusions and sensitivity analyses. The full Study 1 data set is housed on secured university servers, to which only authorized research personnel have password-protected access. Data are organized according to arbitrary participant numbers; identifiable information is stored separately and never publicly shared. Protocols for data collection and storage, including privacy protection, were approved by the Northeastern University Institutional Review Board (IRB# 16-01-13).

Participants. An initial 67 adults were recruited from Northeastern University classrooms and online portals as well as the greater Boston area through posted advertisements; all eligible participants were fluent English speakers. Informed consent was obtained from all participants before beginning the study. Participants received US$490 for completing all parts of the study, plus up to US$55 in compliance and task incentives. Six participants withdrew, nine were dismissed due to poor compliance, and two were excluded from data analysis because they did not complete the full study protocol, for a final sample size of 50 (54% female; 40% White, 2% Black, 44% Asian, 14% other; M = 22.5 years, standard deviation [SD] = 4.4 years). We conducted a sensitivity analysis for this data set in G*Power (Faul et al., 2009), assuming $\alpha < .05$, two-tailed and power (1-β) > .80, which indicated that this data set was adequately powered to detect between-person bivariate correlations of $r \geqslant .38$. A separate sensitivity analysis conducted in the same way indicated we were powered to detect within-person effects of $r \geqslant .34$ to $r \geqslant .21$, depending on the specific number of prompts completed by a given participant.

Procedure. Participants completed approximately 14 days (M = 14.4, SD = 0.6) of experience sampling including peripheral physiological monitoring and end-of-day diaries. Each day of experience sampling lasted for 8 hours and began when participants were outfitted with physiological sensors and a smartphone with an associated smartphone application. Specifically, participants wore a mobile impedance cardiograph that was used to measure the electrocardiogram, impedance cardiogram, and electrodermal activity via wired sensors, as well as movement via three-axis accelerometers. Participants also wore two inertial measurement units to assess posture. Physiological and accelerometric data were recorded continuously throughout the 8-hour sampling period each day and communicated via Bluetooth to a linked smartphone. Most experience sampling prompts were physiologically triggered to enable more efficient sampling of psychologically salient moments (Hoemann et al., 2020). These prompts occurred any time there was a substantial, sustained change in cardiac activity in the absence of movement. Participants also received two "random" prompts each day that were not contingent on changes in cardiac activity. Altogether, participants responded to an average of 8.65 (SD = 1.09) prompts per day or between 63 and 183 prompts overall.

At each sampling prompt, participants responded to a series of questions presented on the smartphone app. First, participants provided a very brief, free-text description of what was going on at the time they received the prompt, intended as a mnemonic for later recall (e.g., "eating lunch with friends"). Next, they rated their current valence on a 100-point continuous slider scale ranging from −50 (very unpleasant) to + 50 (very pleasant). Participants then self-generated words to label their current affective experience. Specifically, participants were asked to "list any emotion(s) you were feeling when you received the prompt." Participants were able to provide as many words as they felt necessary but were required to input at least one word, entering each into a separate, short-response field (15–20 visible characters). For each entry, participants were asked to provide an intensity rating on a Likert-type-style scale from 1 ("not at all") to 5 ("very much"). Participants also responded to additional questions not analyzed in this study.

Upon finishing each day of experience sampling, participants automatically received an online end-of-day diary. In this diary, for each completed sampling prompt from that day, they were presented with the prompt time and the brief description they provided earlier, which served as a guide for participants to provide additional details about their experience at the time of each prompt. As part of this process, participants rated the intensity of their emotional experience on a set of 18 emotion adjectives using Likert-type-style scales from 0 ("not at all") to 6 ("very much"): "afraid," "amused," "angry," "bored," "calm," "disgusted," "embarrassed," "excited," "frustrated," "grateful," "happy," "neutral," "proud," "relieved," "sad," "serene," "surprised," and "worn out." These emotions were selected to sample high-, mid-, and low-arousal octants of the affective circumplex (see Barrett, 1998) using common English adjectives. Intensity ratings were requested in the end-of-day diary, rather than at each experience sampling prompt, to reduce participant burden in the moment.

Participants also attended two in-laboratory sessions, one prior to experience sampling and one after. Among other questionnaires and tasks, they completed the following measures of mental health and emotional functioning at each session: anxiety (Generalized Anxiety Disorder [GAD7]; Spitzer et al., 2006), depression (Patient Health Questionnaire, Depression scale [PHQ-8]; Kroenke et al., 2009), alexithymia (Toronto Alexithymia Scale, 20-item version [TAS-20]; Bagby et al., 1994), and emotional complexity (Range and Differentiation of Emotional Experience Scale [RDEES]; Kang & Shaver, 2004).

Data Preparation. Anonymized data, preparation code, and step-by-step instructions are available via a repository hosted by the Center for Open Science Framework (OSF) at https://osf.io/urt3x/.

Inventorying Emotion Labels. We began by summarizing and describing the set of emotion words used by participants to label their experience at each prompt. To do this, we created a series of scripts in Python 3.9.1. An initial script converted labels to lowercase and tokenized them (i.e., split strings of text into individual words), producing a list of all unique words used by all participants and the number of times they were used. This list was then used to manually create a dictionary of "emotion" words. Adverbs and other modifiers (e.g., "very" and "getting") and function words (e.g., "and" and "in") were not included in these dictionaries. Verbs (e.g., "studying," "relaxing," and "thinking") and nonverbal expressions (e.g., "eww") were retained. These decisions were made to be as liberal as possible in which words were counted, resulting in the inclusion of 429 words. This dictionary was then used to identify the specific words used at each prompt. Words were stemmed prior to comparison using the National Language Toolkit (NLTK; toolbox (Bird et al., 2009). Stemming was used to deal with cases where participants used more than one part of speech to refer to the same emotion (e.g., "happy" and "happiness" should not be treated as unique references). The resulting data were then passed to a final script in MATLAB (2018) for calculating the number of words used per prompt, the total number of words used per person, and the average number per prompt, as well as the number and proportion of the latter that were unique. Prompts where no emotion words could be counted were considered missing data and dropped prior to analysis (55 of 5440, or 1% of prompts), as natural language cannot be imputed.

Recapitulating Valence Ratings. Participants' valence ratings were extracted for each prompt and also averaged per person across all prompts; there were no missing data, and so no form of imputation was used. We used these valence ratings as the criterion for assessing how well momentary valence was captured by freely generated emotion labels. Our next step was to estimate the valence of the (set of) emotion word(s) used at each prompt. In approaching this task, we first considered but decided against several alternative approaches. Many existing sentiment analysis algorithms (see Hartmann et al., 2023) are created to classify texts (e.g., positive vs. negative) and to estimate the confidence or (unipolar) intensity of that classification, rather than to assign a continuous (bipolar) estimate of valence. Word-counting programs such as LIWC and databases of valence norms (see Warriner et al., 2013) cannot account for negation (e.g., "not happy") or other contextual aspects of word use and are not guaranteed to include all the emotion words used by our participants (e.g., "blithe"). To address these gaps, we estimated valence for each prompt by creating a custom Python pipeline as follows.

Specifically, we encoded the National Research Council Canada (NRC) valence norms (Mohammad & Turney, 2013) with the (English) Sentence Bert Python package (Reimers & Gurevych, 2019) so that each of the ˜20,000 words represented by the NRC norms was assigned a different (768-dimensional) vector position (following Song et al., 2020). We used these positions to determine the nearest valence-rated neighbors for each of the "target" words used by our participants (following, see Hollis & Westbury, 2016; see also Di Natale et al., 2021; Recchia & Louwerse, 2015; Van Rensbergen et al., 2016). Target words were appended to the phrase "I feel ___" to provide context, and each phrase was assigned a vector position. Estimated valence was then calculated by taking a weighted average of the 10 nearest neighbors (using cosine distance), with weights inversely proportional to distance. In keeping with the NRC norms database, valence estimates ranged continuously between 0 (extremely unpleasant) and 1 (extremely pleasant). To test the validity of our approach, we compared the valence of 500 randomly sampled words in the NRC database with their estimates generated by our pipeline (after removing the target word from the list of potential neighbors) and found our valence estimation and the ground-truth NRC norms to be highly correlated at r = .89. For the present analyses, if a target word in our sample matched an NRC word, we performed a simple lookup instead of using the estimation pipeline. Though data were not cleaned for typos, punctuation, etc. prior to analysis, the vector space we employed was trained to predict subsequent subword letter groupings rather than whole words (i.e., "byte-pair encoding"). If a word is spelled almost correctly (e.g., "happt"), the subword tokens of the misspelled word will substantially overlap with the subword tokens in the correctly spelled word (i.e., "happy"). The ensuing semantic vectors will be similar, partially obviating the need to correct minor typos.

As a robustness check, we also estimated each prompt's valence more coarsely as the proportion of positive words. To do this, we submitted the list of emotion words identified above to our valence estimation script and used the resulting values to create separate dictionaries of positive (valence $\geqslant$ .5) and negative (valence $<$ .5), for a total of 226 positive and 203 negative words. Subsequent scripts used these dictionaries to count the specific words used at each prompt: a Python script preprocessed and compared prompt text to the word stems in each dictionary; a MATLAB script used these word counts to calculate the proportion of positive words at each prompt and the average of this measure per person. The proportion of negative words was not calculated, as it is the simple inverse of the proportion of positive words.

Testing Convergent Validity. Finally, we compiled measures of mental health and emotional functioning using data from both questionnaires and end-of-day diaries. For mental health, we used the anxiety (GAD-7) and depression (PHQ-8) symptom inventories collected in the second in-laboratory session, as these asked participants to report on the period covered by experience sampling (i.e., the previous two weeks). We observed a strong positive intercorrelation between the GAD-7 and PHQ-8 scores (r = .71), consistent with prior studies that have demonstrated robust associations between self-reported anxiety and depression symptoms (see Clark & Watson, 1991; Feldman, 1993). For this reason, we standardized and then averaged the scores to achieve a single estimate of mental ill-health per participant, where higher scores are associated with worse mental health. To examine whether mental ill-health was differentially associated with the use of unique negative versus positive words (Vine et al., 2020), we also calculated, per person, the proportion of all positive words that were unique and the proportion of all negative words that were unique.

For emotional functioning, we used the total scores for alexithymia (TAS-20) and emotional complexity (RDEES) collected in the second in-laboratory session. We computed estimates of emotional granularity and emodiversity from the emotion intensity ratings from the end-of-day diaries. Prompts missing a rating for at least one emotion adjective were considered missing data and dropped prior to analysis (49 of 5,440, or 1% of prompts); no imputation was used. Emotional granularity was estimated via an intraclass correlation (ICC) for consistency with averaged raters (i.e., "C-k" method; see Kalokerinos et al., 2019). Higher ICC values reflected lower emotional granularity (i.e., greater shared variance among adjectives' ratings). We computed granularity separately for 10 positive emotions ("amused," "calm," "excited," "grateful," "happy," "neutral," "proud," "relieved," "serene," and "surprised") and 8 negative emotions ("afraid," "angry," "bored," "disgusted," "embarrassed," "frustrated," "sad," and "worn out"), with these assignments based on a median split of normative ratings (Warriner et al., 2013). We then averaged these values to create a single estimate of emotional granularity per participant (following, see Edwards & Wupperman, 2017). This two-step procedure avoided interpretation issues that arise from including ratings for all emotion terms in a single ICC; because ratings for positive and negative emotions are typically negatively correlated, including all emotions in the same analysis can result in negative ICC values. ICCs were Fisher r-to-z transformed to fit the variable to a normal probability distribution. These transformed values were multiplied by -1 to yield estimates of granularity that scaled intuitively, such that lower (more negative) values reflected lower granularity, and higher (less negative) values reflected higher granularity. Emotional granularity was calculated using ICC (Salarian, 2016) in MATLAB.

Emodiversity was estimated via a Gini coefficient. This coefficient captures how evenly a phenomenon is observed across various types or categories; here, it captured the relative spread of prompts across the 18 sampled emotion adjectives. Emodiversity was calculated using the formula from Benson et al. (2018) as a custom function in MATLAB.

To summarize, three prompt-level and eight person-level measures were derived from self-generated emotion labels. One prompt-level and one person-level measures were derived from momentary valence ratings. Five additional person-level measures were derived from questionnaires and emotion intensity ratings. These 18 measures are summarized in Table 1.

Analysis. We conducted both descriptive and inferential analyses to accomplish our three study goals. Inferential

Table 1. Measures Used in Studies 1 and 2.

Name	Level	Description	Study 1	Study 2
Derived from free labels				
Number of words	Prompt	Number of (emotion) words used at a given prompt	x	x
Total words	Person	Total number of words used	x	x
Words per prompt	Person	Total number of words divided by the number of prompts completed	x	x
Unique words	Person	Total number of unique words used	x	x
Proportion of unique words	Person	Number of unique words divided by the total number of words	x	x
Estimated valence	Prompt	Valence estimated for all words at a given prompt (0 to 1 continuous; 0 = extremely negative and 1 = extremely positive)	x	x
Average estimated valence	Person	Mean estimated valence across all prompts	x	x
Proportion of positive words	Prompt	Number of words with valence ⩾ .5 divided by number of words	x	x
Average proportion of positive words	Person	Mean proportion positive words across all prompts	x	x
Proportion of unique positive words	Person	Number of unique words with valence ⩾.5 divided by the total number of words with valence ⩾ .5	x	x
Proportion of unique negative words	Person	Number of unique words with valence < .5 divided by the total number of words with valence < .5	x	x
Self-reported valence	Prompt	Continuous valence reported at a given prompt (−50 to + 50 continuous; −50 = extremely negative and + 50 = extremely positive)	x	x
Average self-reported valence	Person	Mean self-reported valence across prompts	x	x
Emotional granularity	Person	Variance shared among emotion adjectives (ICC)	x	
Emodiversity	Person	Evenness of distribution across emotion adjectives (Gini coefficient)	x	
Derived from questionnaires				
Mental ill-health	Person	Averaged standardized anxiety (e.g., GAD-7), depression (e.g., PHQ-8) scores	x	x
Alexithymia	Person	TAS-20 total score	x	
Emotional complexity	Person	RDEES total score	x	

analyses were conducted with two-tailed significance at $\alpha = .05$ and outliers retained. Linear mixed-effects models were fitted with *lme4* (Bates et al., 2015) in R (R Core Team, 2020) on within-sample standardized variables. Remaining analyses were performed in MATLAB. Analytic code is available via our OSF repository.

Inventorying Emotion Labels. We quantified the number of unique words used across the sample and visualized the most frequent using word clouds. Word clouds were generated with raw individual words, prior to stemming, to aid readability. We qualitatively interpreted the set of words as a whole, looking in particular for labels that diverged from those typically used to gather structured ratings, as well as for particularly nuanced or rich descriptions that provided additional insight into participants' momentary experience. Finally, we verbally and visually described the number of words used per prompt, the total words per person and average per prompt, as well as the number and proportion of the latter that were unique.

Recapitulating Valence Ratings. We first tested the relationship between average estimated valence and average self-reported valence using a bivariate correlation. We then tested it within persons as the correlation between estimated valence and self-reported valence at the prompt level separately per participant. We also fitted a linear mixed-effects regression predicting self-reported valence from estimated valence, in which intercepts and slopes were free to vary per participant. We repeated these steps using the proportion of positive words instead of estimated valence to assess the impact of continuous versus count-based measures.

Testing Convergent Validity. To examine whether the valence and uniqueness of emotion labels were related to self-reported anxiety and depression, we correlated average estimated valence, average proportion of positive words, and proportions of unique positive/negative words between persons with mental ill-health.

To examine whether the number and uniqueness of emotion labels were related to self-reported and derived measures of emotional functioning, we correlated words per prompt and the proportion of unique words between persons with alexithymia, emotional complexity, emodiversity, and emotional granularity.

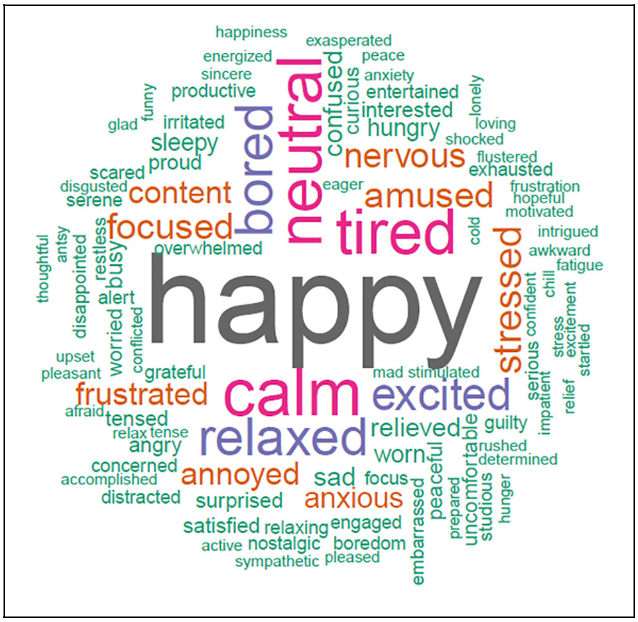

Figure 1. Word Cloud of the 100 Most Frequently Used Emotion Labels, Study 1.

Results

Inventorying Emotion Labels. Altogether, participants used 429 unique words to label their current feelings (see our OSF repository for a complete list). A word cloud of the 100 most common emotion labels is presented in Figure 1. Descriptive statistics for measures derived from these labels are presented in Table 2.

We observed that participants generated between 1 and 5 emotion words per prompt. "Happy" was the most frequent label (used 1,213 times); many typically sampled emotions were also used regularly (e.g., "relaxed," "excited," and "bored" >300 times each), although some were not (e.g., "disgusted" only 8 times). Labels often referenced body-focused states such as

Table 2. Descriptive Statistics for Inventorying Emotion Labels, Study 1.

Measure	Level	*M*	*SD*	Minimum	Maximum
Number of words	Prompt	1.51	.55	1	5
Total words	Person	164.28	76.87	67	447
Words per prompt	Person	1.51	.55	1	3.04
Unique words	Person	33.08	23.68	7	139
Proportion of unique words	Person	.20	.09	.07	.46

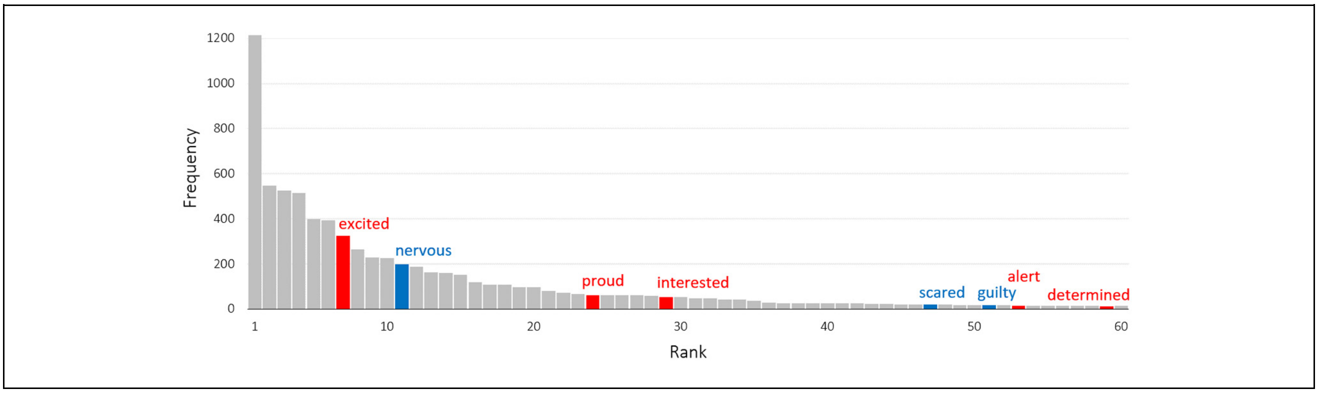

Figure 2. The 60 Most-Frequently Used Emotion Labels in Study 1 and Where the PANAS Words Are in the Frequency Ranking (Highlighted in Red and Blue, Respectively, for Positive Affect and Negative Affect; Y. Li et al., 2020).

"tired" and "hungry," cognition-focused states such as "focused" and "confused," and activity-focused states such as "busy" and "lazy" that are not common in studies of emotional experience. Other noteworthy labels included "uncomfortable," "overwhelmed," "productive," "nostalgic," "conflicted," "awkward," "determined," "lonely," "impressed," "antsy," "chill," "defeated," "appreciative," "optimistic," "chagrined," "challenged," "resigned," "unconcerned," "stoked," and "blithe." Participants also made use of short phrases such as "worn out," "fired up," "mentally exhausted," "getting stressed," "at ease," "taken aback," "in pain," and "creeped out."

Following Y. Li et al. (2020), we compared the words generated by participants to those sampled in the 20-item PANAS (Watson et al., 1988). As can be seen in Figure 2, few PANAS terms were among the most frequently used labels, replicating the findings from Y. Li et al. (2020) and suggesting that contemporary, daily life emotions do not often map onto those in established scales. Seven PANAS items were generated fewer than 10 times total; four were not generated at all. See Supplemental Table S1 for the full comparison, including the terms sampled by the 60-item PANAS-X (Watson & Clark, 1994).

Participants used as few as 7 and as many as 139 unique words across experience sampling. The mean number and proportion of unique words were 33.08 and .20, respectively, meaning that on average participants established a medium set of labels and returned to it over time, though this varied considerably across individuals.

Recapitulating Valence Ratings. Between persons, average estimated valence was positively associated with average self-reported valence ($r = .66$, $p < .001$). Participants who tended to provide higher valence ratings in the moment also tended to label their current feelings using-with words with stronger positive connotations. Within persons, the association between estimated and self-reported valence was positive on average ($Mr = .52$; 48 of 50 $p < .05$), but the strength of this association varied considerably across participants ($SDr = .18$; range: $-.03 < r < .79$). A linear mixed-effects model confirmed that the prompt-level relationship was positive and significant overall when accounting for participant-level variation, $b = .49$, standard error $(SE) = .04$, $t(48.30) = 13.62, p < .001$.

A parallel set of analyses using the proportion of positive words instead of estimated valence revealed that the count-based measure achieved virtually identical results to the continuous one. The between-persons correlation was $r = .60$, $p < .001$, with within-persons correlations averaging $Mr = .47$ (44 of 50 $p < .05$) and varying considerably across participants ($SDr = .20$; range: $-.05 < r < .78$). In a linear mixed-effects model, proportion of positive words per prompt significantly predicted self-reported valence per prompt accounting for participant-level variation, $b = .47$, $SE = .04$, $t(48.03) = 11.11, p < .001$.

Testing Convergent Validity

Relationships With Mental Ill-Health. Between persons, average estimated valence was negatively associated with self-reported mental ill-health ($r = -.40$, $p = .01$). Participants who tended to label their current feelings using words with stronger positive connotations during experience sampling also reported fewer symptoms of anxiety and depression when reflecting on the experience sampling period. We observed a similar association between the average proportion of positive words and mental ill-health ($r = -.32$, $p = .03$). Participants who used relatively more positive words reported fewer symptoms of anxiety and depression. Note that we were slightly underpowered to detect this second effect (sensitivity analysis suggested appropriate power to detect $r \geqslant .38$), and both were descriptively smaller though similar in magnitude to the association between average self-reported valence and mental ill-health ($r = -.45$, $p < .001$). There was no association between mental ill-health and the overall proportion of unique negative words ($r = -.24$, $p = .09$) nor the overall proportion of unique positive words ($r = .14$, $p = .32$).

Relationships With Emotional Functioning. Between persons, the average number of words per prompt was not associated with self-report measures of alexithymia ($r = -.10$, $p = .51$) or emotional complexity ($r = .13$, $p = .36$), nor was it associated with measures of emodiversity ($r = .25$, $p = .09$) or emotional granularity ($r = .15$, $p = .30$) derived from the emotion intensity ratings in the daily diaries. The overall proportion of unique words was also unrelated to alexithymia ($r = .07$, $p = .64$), emotional complexity ($r = .06$, $p = .66$), and emotional granularity ($r = -.19$, $p = .19$). However, it was positively associated with emodiversity ($r = .47$, $p < .001$). Participants who repeated emotion labels less often in the moment endorsed a more even spread of emotions when reflecting on their experiences at end of day.

In ancillary analyses, we examined whether separate estimates of emotional granularity and emodiversity for negative and positive emotions were related to the overall number and uniqueness of (negative/positive) emotion labels. These analyses revealed null associations, with two exceptions: The average number of positive words per prompt was positively associated with positive emotional granularity, and the overall proportion of unique positive words was positively associated with positive emodiversity. For details, see Pages 2 to 3 of the Supplemental Materials.

Discussion

In Study 1, participants provided self-generated emotion labels and continuous valence ratings to describe how they were currently feeling at each experience sampling prompt for 14 days, provided retrospective emotion intensity ratings for each prompt in end-of-day diaries, and reported on their anxiety/depression symptoms and emotional functioning in post-sampling questionnaires. Secondary analysis of this data set revealed extensive variation in the words and short phrases participants used to describe their experiences. These self-generated emotion labels recapitulated rated valence both between and within participants, and measures of emotion label valence were associated with mental ill-health. A measure of emotion label uniqueness was also associated with emotional functioning in the form of emodiversity. To examine whether these effects were specific to this data set, and to address issues of low statistical power, in Study 2 we sought to replicate a subset of these findings in a data set including momentary self-generated emotion labels and valence ratings, as well as questionnaire measures of mental health.

Study 2

Methods

Study 2 data are reported in two other publications (MacVittie et al., under review and MacVittie et al., in press), though these have not included the open-ended emotion labels. Below we report aspects of the methods relevant for the present analyses, including all data exclusions and sensitivity analyses. The full Study 2 data set is housed on secured university servers, to which only authorized research personnel have password-protected access. Data are organized according to arbitrary participant numbers; no identifiable information was collected. Protocols for data collection and storage were approved by the University of New Hampshire Institutional Review Board (IRB# IRB-FY2022-162). Analyses for Study 2 were preregistered at https://osf.io/xceu5.

Participants. Study 2 followed an opt-in remuneration protocol: Participants completed an initial set of surveys and were then asked to enroll in the experience sampling portion. An initial 160 adults were recruited through Prolific, an online participant pool in which people can volunteer to sign up for experiments and surveys. Eligible participants were fluent English speakers who were able to complete a 28-day study. Participants also needed daily access to a smartphone and network connectivity required for experience sampling. Informed consent was obtained from participants before beginning the study. Participants were compensated up to US$40 for completing all surveys and up to US$42 for completing all experience sampling prompts; participants who completed at least 75% of the prompts received a US$10 bonus at the end of the study. An additional US$20 in compensation and incentives came from tasks not included in the present analyses, for a possible total of US$112. Forty-two participants did not enroll in the experience sampling portion and, as described below, nine were excluded from data analysis because they completed too few experience sampling prompts. Of the remaining 109 participants, a further 12 did not complete the survey used in the present analyses. The final sample size was 97 (44% female; 55% White, 26% Black, 1% Asian, 13% other; $M = 28.4$ years, *SD* = 8.9 years). A sensitivity analysis conducted as above indicated that this data set was adequately powered to detect between-person effects of $r \geqslant .28$.

Procedure. Participants completed a 28-day (4-week) experience sampling protocol in which they received six prompts per day via a dedicated app on their smartphone; prompts were sent at pseudorandom times between 10 am and 8 pm and at least 1 hour apart. The total number of completed prompts ranged between 2 and (the full) 168 per person. To ensure we were adequately powered to detect the within-person effect sizes from Study 1 ($M_r \geqslant .50$), we conducted an a priori power analysis for bivariate correlations assuming $\alpha < .05$ (two-tailed) and $1\text{-}\beta > .80$. This analysis suggested a minimum (within-person) sample size of 29. Accordingly, we excluded the abovementioned participants who each completed fewer than 30 prompts.

At each sampling prompt, participants responded to a series of questions, including to rate their current valence on Likert-type-style scales ranging from 1 (extremely unpleasant) to 7 (extremely pleasant). They were also asked to provide a word or phrase that described their current feeling via an open text box.

Prior to experience sampling, participants provided basic demographic information and completed measures of depression (via the Patient-Reported Outcomes Measurement Information System Depression Disorder Questionnaire, i.e., PROMIS Depression; Pilkonis et al., 2011) and anxiety (via the Patient-Reported Outcomes Measurement Information System Anxiety Disorder Questionnaire, i.e., PROMIS Anxiety; Pilkonis et al., 2011). Participants repeated these mental health measures twice during experience sampling: once after 2 weeks (in the middle of the study) and again after 4 weeks (at the end of the study). All other data collected in the study were excluded from the present preregistration and analyses.

Data Preparation. Data preparation proceeded as described for Study 1, with a few notable exceptions. Because no emotion intensity ratings were collected, we could not derive measures of emotional granularity or emodiversity. Similarly, no self-report measures of alexithymia or emotional complexity were collected. We once again used the self-report measures of anxiety and depression collected after experience sampling. These were highly correlated ($r = .84$) and so were standardized and averaged to achieve a single estimate of mental ill-health per participant. Prompts with blank entries (4 of 13,553, or 0.03%) or where no emotion words could be counted (592 of 13,549, or 4.37%) were dropped prior to analysis. There were no missing valence ratings. Our data-driven word-counting dictionaries retained 746 words (344 positive and 402 negative).

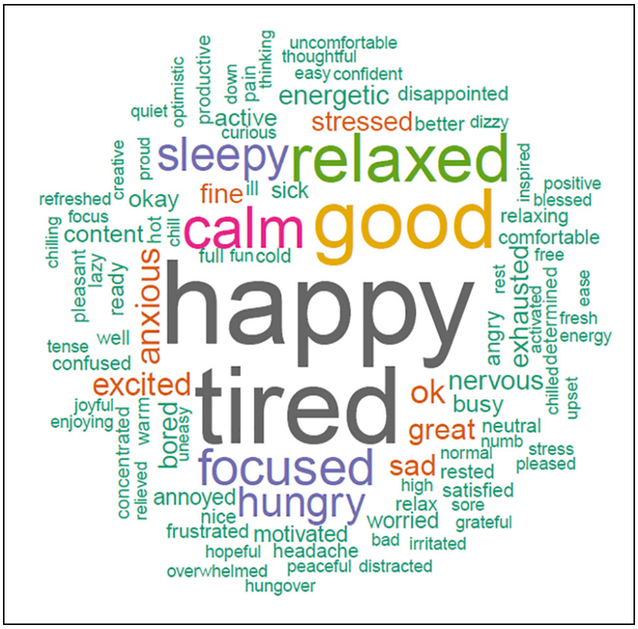

Figure 3. Word Cloud of the 100 Most Frequently Used Emotion Labels, Study 2.

Analysis. Analyses were conducted as described for Study 1. As preregistered, we focused on the between- and within-persons relationships between estimated valence and self-reported valence, and on the between-persons relationships of estimated valence and the proportions of unique positive/negative words with mental ill-health. We also conducted exploratory analyses not included in our preregistration, in which we examined the between- and within-person relationships between the proportion of positive words and self-reported valence, as well as the between-persons relationship of the mean proportion of positive words with mental ill-health.

Results

Inventorying Emotion Labels. Study 2 participants used 746 unique words to label their current feelings (see our OSF repository for the complete list). A word cloud of the 100 most common emotion labels is presented in Figure 3. As in Study 1, the word cloud was generated with raw individual words, prior to stemming, to aid readability. Descriptive statistics for measures derived from these labels are presented in Table 3.

Participants in Study 2 generated between 1 and 11 emotion words per prompt. In contrast to Study 1,

Table 3. Descriptive Statistics for Inventorying Emotion Labels, Study 2.

Measure	Level	*M*	*SD*	Minimum	Maximum
Number of words	Prompt	1.62	.73	1	11
Total words	Person	220.95	121.27	38	866
Words per prompt	Person	1.62	.73	1	5.38
Unique words	Person	46.55	27.79	6	138
Proportion of unique words	Person	.23	.10	.02	.45

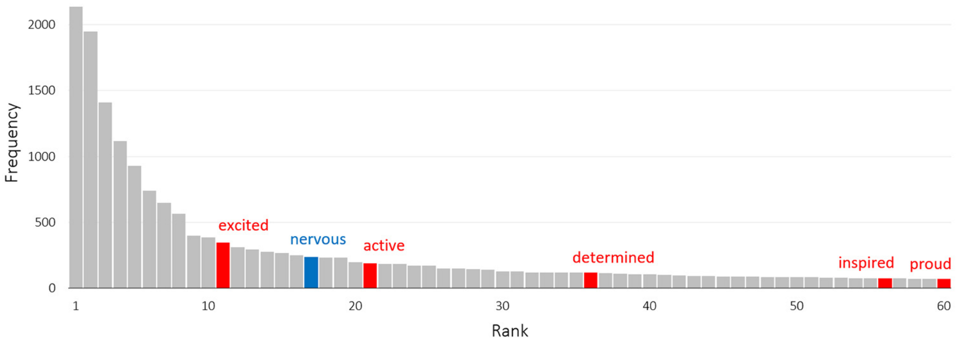

Figure 4. The 60 Most-Frequently Used Emotion Labels in Study 2 and Where the PANAS Words Are in the Frequency Ranking (Highlighted in Red and Blue, Respectively, for Positive Affect and Negative Affect; Y. Li et al., 2020).

where responses were limited to short phrases (e.g., "worn out"), some Study 2 participants chose to write complete sentences; total response length varied between 1 and 52 words ($M = 3.32$, $SD = 3.93$). In these longer responses, participants usually provided the context or reason for their current feelings, for example: "I feel happy as I got enough sleep last night"; "I'm sad, 7 years [ago] my best friend died of cancer"; "Thrilled because my brother is coming to visit me for my birthday"; "I am frustrated because I want to study but I have a headache"; "Sad, shocked, appalled by the assault story I am reading on social media"; and "I am feeling so so excited, I got a beautiful day planned ahead, I just cannot wait."

"Happy" was again the most frequent label (used 2,135 times), with other typically sampled emotions distributed similarly to Study 1 (e.g., "relaxed" used 1,116 times and "disgusted" 7 times). "Tired" was the second-most frequent label (1,949 times), with "sleepy," "hungry," "sick," and other body-focused states also near the top of the list. Cognition-focused states (e.g., "focused"; 739 times) and activity-focused states (e.g., "busy"; 196 times) were likewise popular. In comparison with Study 1, we noticed heavier usage of generic or nonspecific, "affective" labels such as "good," "ok(ay)," "great," "fine," and "bad." Unique labels from Study 2 included "numb," "sentimental," "resentful," "somber," "moody," "courageous," "gleeful," "homesick," "fragile," "powerless," "demotivated," "cringe," "pessimistic," "vibing," "Zen," "desperate," "unstoppable," "compassionate," "rejuvenated," and "appalled."

Comparing the words generated by participants in Study 2 to those sampled in the PANAS, we once again found that few PANAS terms were among the most frequently used labels (Figure 4). Six PANAS items were generated fewer than 10 times total; one was not generated at all. See Supplemental Table S2 for the full comparison.

The number of unique words used per person ranged between 6 and 138, with an average of 46.55. As in Study 1, the mean proportion of unique words, .23, suggested that participants on average reused a moderate number of labels to describe their everyday emotional experiences, though this again varied considerably across participants.

Recapitulating Valence Ratings. Replicating Study 1, average estimated valence was positively and significantly associated with average self-reported valence in a between-persons analysis, $r = .79, p < .001$. Within persons, the strength of the association between estimated and self-reported valence at the prompt level was again positive on average ($Mr = .46$; 83 of 97 $p < .05$) but varied in strength across participants ($SDr = .19$; range: $.02 < r < .81$). A linear mixed-effects model confirmed that the prompt-level relationship was positive and significant when accounting for participant-level variation, $b = .44$, $SE = .02$, $t(93.22) = 22.41$, $p < .001$.

Exploratory analyses using the proportion of positive words instead of estimated valence again revealed similar results for the count-based measure. The between-persons correlation was $r = .68$, $p < .001$, with within-persons correlations averaging $Mr = .40$ (80 of 97 $p < .05$) and varying considerably across participants ($SDr = .19$; range: $-.14 < r < .76$). The overall relationship was also confirmed by a linear mixed-effects model accounting for participant-level variation, $b = .39$, $SE = .02$, $t(91.49) = 18.02$, $p < .001$.

Testing Convergent Validity: Relationships With Mental Ill-Health. Replicating Study 1, average estimated valence was negatively and significantly associated with self-reported mental ill-health, $r = -.62$, $p < .001$. In an exploratory analysis, we observed a virtually identical association between the average proportion of positive words and mental ill-health, $r = -.62$, $p < .001$. These associations were of the same strength as that between self-reported valence and mental ill-health ($r = -.63$, $p < .001$). Like in Study 1, the overall proportion of unique negative words was not associated with mental ill-health, $r = -.11$, $p = .30$. Unlike in Study 1, there was a strong positive association between the overall proportion of unique positive words and mental ill-health ($r = .55$, $p < .001$), such that participants who used a relatively larger number of unique, positively valenced words also reported more symptoms of anxiety and depression.

Discussion

In Study 2, participants provided self-generated emotion labels and continuous valence ratings to describe how they were currently feeling at each experience sampling prompt for 28 days and reported on their anxiety/depression symptoms in post-sampling questionnaires. As in Study 1, participants used a diverse range of words and short phrases (and occasionally even full sentences) to relate their experiences. Preregistered and adequately powered analyses again showed that these self-generated emotion labels recapitulated rated valence both between and within participants, and that measures of emotion label valence were associated with mental ill-health.

Measures of label uniqueness were associated with mental ill-health differently than they were in Study 1. Namely, participants who used more unique words for positive emotion over the course of experience sampling also reported more symptoms of anxiety and depression. This finding is broadly in keeping with work showing that reporting on pleasant experiences with greater precision inhibits appreciation and enjoyment (i.e., savoring; see Starr et al., 2017). It is possible that diversity of emotion word use reflects regular attempts to identify the "correct" label; these attempts may indicate (or induce) a more abstract or psychologically distanced mode of processing, which in turn may signal dampening or avoidance of positive emotions. Future work could test this possibility by examining the semantic complexity of the emotion words generated, or by assessing positive emotion regulation strategies (see Quoidbach et al., 2010).

General Discussion

Day-to-day interactions and moments of reflection do not typically involve evaluating feelings along a standard list of emotions ("2 out of 5 happy," "1 out of 5 calm," "4 out of 5 excited," etc.) or pleasantness on a scale of 0 to 100. Instead, people are more likely to say, "tired but happy," "focused on my work," or "so excited I cannot wait." Despite this intuition, experience sampling studies often adopt the former, structured approach to assessing emotion, potentially losing the subtlety and diverseness of the original feelings in the process. Across two studies, we examined what could be learned by allowing participants to describe their current experiences in their own words. We found a wide range of labels for feelings that went far beyond those typically sampled in psychological studies. Participants varied greatly in which labels they used and how they used them. Nonetheless, estimates of label positivity were consistent with momentary valence ratings both within and across participants, and were associated with fewer self-reported symptoms of anxiety and depression. Using additional measures available in the first study, we found that the number of unique emotion words used over time was related to the balance and spread of emotions endorsed in an end-of-day rating task (i.e., emodiversity), but not to other measures of emotional functioning (e.g., alexithymia). Considering the sheer interpretative value of the raw participant responses, a free-label approach appears to be a viable and compelling means of studying emotion in everyday life.

The present research applied a free-generation approach to assessing self-reported emotion (Y. Li et al., 2020) in the context of daily life. While ours is the first work we are aware of that asked participants to label their current momentary experiences (but see Ottenstein & Lischetzke, 2019 for a daily approach), its findings build on several recent studies that have assessed emotion using open-ended, text-based methods. In the work by Carlier et al. (2021), experience sampling participants optionally recorded descriptions of what they were doing and how they felt about it at each prompt.

Estimates of positive and negative affects derived from the transcribed recordings (using LIWC; Pennebaker et al., 2015) were very modestly associated with participant-reported valence both within and between persons, but performed better than estimates from acoustic analyses or concurrent text messages. This study, like the present, contrasts with other work that reports no or limited correspondence between momentary ratings and estimates derived from unobtrusively gathered language (e.g., from recordings of ambient speech; Sun et al., 2019). Our studies demonstrate that when descriptions of current experience are explicitly requested and deliberately provided, they can give insight into affective dynamics and other relevant indicators. Compared with studies that have asked participants to describe their experiences at length (see Carlier et al., 2021), however, we show that even one or a few words is sufficient to recover reliable valence estimates and predict mental health and emotional functioning over time.

Emotion labels are not limited to estimates of valence, but also enable qualitative and quantitative comparison of the emotion word repertoires in use. The past few years have seen multiple initiatives for assessing and comparing spontaneously generated emotion words. In one approach, estimates of emotional granularity are generated by coding the emotion words used to describe daily experiences for their level of specificity (Ottenstein & Lischetzke, 2019; see also Thompson et al., 2021; Williams & Uliaszek, 2021). This specificity index, like the measures of word use in the present studies, showed some associations with mental health symptoms but not with the traditional, rating-based estimate of emotional granularity. In another approach, estimates of emotion word diversity are generated by counting the number of unique emotion concepts invoked (e.g., using "happy" or "happiness" for *happy*) as a function of text length (Vine et al., 2020). The diversity of positively and negatively valenced words in stream-of-consciousness essays and blogs were differentially associated with health and distress, with positive word diversity linked to better outcomes and negative word diversity to worse (see also Entwistle et al., 2023). We did not conceptually replicate these findings with our measures of unique positive and negative words; in fact, in Study 2, we found positive word uniqueness to be linked to worse mental health. These discrepancies suggest that the relationship between emotion word repertoires and pertinent outcomes varies based on the context of language use. Indeed, recent work has shown that measures of emotional functioning and well-being are unrelated to fluency for emotion words, when this is assessed in a task (Hegefeld et al., 2023). Future work is necessary to probe how and when using (more) words for emotion is beneficial, and how and when it is not.

A related task for future work is to investigate if and how the number and type of emotion labels generated maps onto the experience of mixed feelings. It is an open question if people providing more than one label at a given time point were describing simultaneous or complex emotions ("happy," "nervous" [Study 1]), or rather using multiple words for the same apparent emotion ("frustrated," "annoyed" [Study 1]). Although we find evidence of both types of responses in our studies, we do not believe our data are ideal for addressing this question. The present studies used a single, bipolar scale to measure and estimate valence. For this reason, we are unable to assess mixed feelings as a combination of positive and negative affects (see J. T. Larsen & McGraw, 2011). Future studies can address this by using separate, unipolar scales and by examining how these measures intersect with instances where multiple labels are provided (or, even, when ambivalent single labels are provided, such as "bittersweet" [Study 2]). To evaluate whether two or more emotion labels are being used synonymously, longer descriptions of momentary experiences that capture situated word meaning are necessary. For example, recent work has used word embeddings at the sentence level to assess the context-specificity of emotion word use (i.e., emotional granularity) in diary entries and consumer reviews over time (Faraji-Rad et al., 2024).

For future work to produce reliable results, it must first come to a standard definition of what emotion words are. That is, what should be considered a reference to an emotion, when used in speech or text? There is a long history of thinking and inquiry on this topic (see Clore et al., 1987; Johnson-Laird & Oatley, 1989; Ortony et al., 1987) but not yet consensus. In the present studies, we took a flexible approach, defining emotion words as any label participants used to describe how they were currently feeling. A benefit of this method is that it allowed us to inventory, in a comprehensive and data-driven manner, the words and phrases used to summarize everyday emotions. These labels did not always match our researcher expectations, shedding light onto the states most salient for our sampled (U.S., English-speaking) population. For example, there was a high prevalence of bodily states like hunger and fatigue, and of cognitive states like focus and confusion. The resulting word inventories echo other recent attempts to create dictionaries of emotion words (see Ottenstein & Lischetzke, 2019). However, our flexible approach also entails the challenge of manually collating, per data set, the emotion words used and may be one reason that our results diverge from prior findings. What "counts" as emotion is likewise culturally shaped (Hoemann, Gendron, et al., 2023; Ip et al., 2023). Moving forward, the field could increase consistency in the identification of emotion words by creating master lists vetted per language, potentially employing distributional semantic models to assess the similarity and uniqueness of word meanings (following Z. Li et al., 2023).

Another, more applied value of our proposed approach may lie in the clinical realm. Experience sampling is increasingly used in blended care, where practitioners prescribe an app to their clients to collect data on the context-specific manifestation of mood, symptoms, and more. These data are then fed back to the practitioner to support diagnostic and therapeutic use, as well as to strengthen client involvement and therapeutic alliance. A significant barrier for widespread uptake of this approach lies in the tension between trying to gather responses that are relevant to and descriptive of each client, while also not overburdening practitioners with customized survey creation and interpretation (Piot et al., 2022; Weermeijer et al., 2023). Adding freely generated labels to describe current feelings could be a means of supplying the client–practitioner relationship with the kind of rich, nuanced, and personally meaningful data that can bolster the effectiveness of treatment, without significantly increasing burden on the side of the client. Critically, our findings show that estimated label valence is equivalently predictive of mental health symptoms, at the between-persons level, as more commonly used valence ratings, suggesting that a free-label approach would not hinder practitioners' ability to detect substantive clinical outcomes. Moreover, it allows practitioners to obtain a picture of fluctuations in clients' mood based on their own words, which can help to bolster therapeutic alliance and ownership.

With all this said, the present work was a first demonstration of a free-label method and pipeline, aiming to showcase its possibilities in a research context. Future research will need to be done that enables automation and benchmarking before a free-label approach can be rolled out at scale, particularly in clinical settings. For example, a better understanding of typical patterns and trends in free-label data is necessary for easily evaluating and assessing participants or clients (e.g., based on the number of emotion words used). At the same time, simpler applications of this approach are already possible or may not be that far off. Free-label responses are supported by virtually all experience sampling apps, many of which are free to use and designed with clinicians in mind (see PsyMate^{TM}, Daniëls et al., 2023; m-Path, Mestdagh et al., 2023). In the near term, it should be feasible for experience sampling apps to integrate valence estimation for free labels, adding to the insight these provide. In the longer term, our property estimation approach could be extended to descriptions of momentary experience that are longer or even spoken (see Atmaja et al., 2020; but see Weidman et al., 2019) or used to capture other meaningful affective or semantic dimensions such as arousal and dominance (Hollis & Westbury, 2016), concreteness or sensorimotor grounding (Altarriba et al., 1999; Wingfield & Connell, 2023), or complexity (Hoemann, Vicaria, et al., 2021; Z. Li et al., 2023). Free labels could also be used as a baseline for constructing personalized lists of items to rate (following Olthof et al., 2023), and this process could be natively supported by apps.

There are additional methodological considerations to keep in mind when evaluating the present findings. The two data sets we used are, to our knowledge, the only ones to use labels instead of specific emotion ratings to capture experiences in everyday life. The samples were also relatively small and homogeneous, comprising people motivated to complete lengthy sampling protocols who may not be representative of the general population on a number of characteristics (e.g., conscientiousness, although this concern pertains to experience sampling studies in general; for discussion, see Wrzus & Neubauer, 2022). Moreover, Study 1 followed an intensive protocol (including physiological recording and end-of-day diaries) that is not representative of other work using experience sampling (as discussed by Hoemann, Barrett, & Quigley, 2021), and Study 2 was not originally designed to test the full set of questions we were able to ask in Study 1. These proof-of-concept demonstrations of the empirical and clinical potential of a free-labeling approach should be replicated in larger, more heterogeneous samples. Such work can also collect measures of overall vocabulary to examine its intersection with the number and uniqueness of emotion words used (see Hegefeld et al., 2023; Suvak et al., 2011), especially if participants need not be native speakers of the language of data collection (as in the present studies). Equally, it is important to note that the present studies were conducted in American English, a known outlier both linguistically and culturally (Blasi et al., 2022; Henrich et al., 2010). Studies in other languages—needed to show whether our findings are generalizable—would need to create their own emotion word inventories and valence estimation pipelines. The latter is imminently possible for major world languages with established valence norms (see Mohammad & Turney, 2013) and pretrained semantic space vectors (see Reimers & Gurevych, 2020).

Asking people how they are feeling is the most natural way of assessing emotional experience and is how people normally communicate with others. The present studies contribute to the field by illustrating the utility of a free-labeling approach for both research and clinical objectives. There are clear pros to this approach: Words provide insight into the rich variation in emotional experience, within and between persons, without limiting participants to, or saddling them with, standard lists of emotion words (Eisele et al., 2020; Y. Li et al., 2020). A free-labeling approach is also language-neutral and can be used across cultures. There are also cons: Some participants may find open-ended response formats more demanding than rating or selecting from a list. Natural language also requires more time and effort for data analysis. Here, we introduced a semi-automated processing pipeline that can reduce burden on research teams interested in implementing this approach but is not yet standardized or scalable to clinical practice. In comparison, structured ratings can be more immediately analyzed and compared by researchers and clinicians alike, as well as used to calculate conventional estimates of emotional granularity and other within- and between-person measures of emotional experience. Structured ratings may also be quicker for participants and clients to complete (especially if momentary valence is the sole interest), and they have the potential therapeutic benefit of expanding awareness of various emotions (see Widdershoven et al., 2019). Ultimately, we see open- and closed-response formats as complementary rather than antagonistic and hope that the advantages and promises of the former, coupled with our initial set of findings, will encourage more work to use labels instead of, or in addition to, ratings to assess emotion in everyday life.

Author Contributions

J.B.W. designed Study 1 with its other principal investigators. K.H. and J.B.W. assisted with data collection for Study 1. J.B.W., C.M., and L.A. designed and supervised data collection for Study 2. K.H. and E.W. analyzed the data. K.H. drafted the article. All authors provided revisions and approved the final version of the article.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: P.K. is shareholder and member of the board of directors of m-Path BV, a KU Leuven spin-off company commercializing m-Path, a platform for experience sampling research and clinical practice. m-Path was not used for data collection in the studies reported in this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: K.H. was supported by the Research Foundation—Flanders (12A3923N). P.K. was supported by the KU Leuven Research Council (C14/23/062, IBOF/21/090). Study 1 was supported by the U.S. Army Research Institute for the Behavioral and Social Sciences (W911NF-16-1-0191 to J.B.W. and Dr. Karen S. Quigley). The views, opinions, and/or findings contained in this article are those of the authors and shall not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents. Study 2 was supported by funds from the Dean's office in the College of Liberal Arts at the University of New Hampshire.

ORCID iD

Katie Hoemann https://orcid.org/0000-0002-9938-7676

Supplemental Material

Supplemental material for this article is available online.

References

Altarriba, J., Bauer, L. M., & Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, and Computers*, *31*(4), 578–602. https://doi.org/10.3758/BF03200738

Atmaja, B. T., Hamada, Y., & Akagi, M. (2020, 16–19 November). Predicting valence and arousal by aggregating acoustic features for acoustic-linguistic information fusion. In *IEEE Region 10 Conference* (pp. 1081–1085). IEEE.

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*(1), 23–32.

Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition and Emotion*, *12*(4), 579–599. https://doi.org/10.1080/026999398379574

Barrett, L. F., Gross, J., Christensen, T. C., & Benvenuto, M. (2001). Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition and Emotion*, *15*(6), 713–724. https://doi.org/10.1080/02699930143000239

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benson, L., Ram, N., Almeida, D. M., Zautra, A. J., & Ong, A. D. (2018). Fusing biodiversity metrics into investigations of daily life: Illustrations and recommendations with emodiversity. *The Journals of Gerontology: Series B*, *73*(1), 75–86. https://doi.org/10.1093/geronb/gbx025

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*, 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Carlier, C., Niemeijer, K., Mestdagh, M., Bauwens, M., Vanbrabant, P., Geurts, L., van Waterschoot, T., & Kuppens, P. (2021). In search of state and trait emotion markers in mobile-sensed language: A field study. *JMIR Mental Health*, *9*, Article e31724. https://doi.org/10.2196/31724

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology*, *100*(3), 316–336.

Clore, G. L., Ortony, A., & Foss, M. A. (1987). The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology*, *53*(4), 751–766. https://doi.org/10.1037/0022-3514.53.4.751

Csikszentmihalyi, M., & Figurski, T. J. (1982). Self-awareness and aversive experience in everyday life. *Journal of Personality*, *50*(1), 15–19. https://doi.org/10.1111/j.1467-6494.1982.tb00742.x

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease*, *175*, 526–536.

Daniëls, N. E. M., Verhagen, S. J. W., van Bokhoven, M. A., Beurskens, A. J., & Delespaul, P. A. E. G. (2023). How to use experience-sampling technology to understand daily functioning: A practical guide for mental health professionals. *Clinical Psychology & Psychotherapy*, *30*(2), 357–372. https://doi.org/10.1002/cpp.2798

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, *3*(5), 478–491. https://doi.org/10.1038/s41562-019-0555-0

Demiralp, E., Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Barrett, L. F., Ellsworth, P. C., Demiralp, M., Hernandez-Garcia, L., Deldin, P. J., Gotlib, I. H., & Jonides, J. (2012). Feeling blue or turquoise? Emotional differentiation in major depressive disorder. *Psychological Science*, *23*(11), 1410–1416. https://doi.org/10.1177/0956797612444903

Diener, E., Larsen, R. J., & Emmons, R. A. (1984). Person × Situation interactions: Choice of situations and congruence response models. *Journal of Personality and Social Psychology*, *47*(3), 580–592. https://doi.org/10.1037/0022-3514.47.3.580

Di Natale, A., Pellert, M., & Garcia, D. (2021). Colexification networks encode affective meaning. *Affective Science*, *2*(2), 99–111. https://doi.org/10/gmhkrg

Edwards, E. R., & Wupperman, P. (2017). Emotion regulation mediates effects of alexithymia and emotion differentiation on impulsive aggressive behavior. *Deviant Behavior*, *38*(10), 1160–1171. https://doi.org/10.1080/01639625.2016.1241066

Eichstaedt, J. C., & Weidman, A. C. (2020). Tracking fluctuations in psychological states using social media language: A case study of weekly emotion. *European Journal of Personality*, *34*(5), 845–858. https://doi.org/10/ghjdzp

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*, 136–151. https://doi.org/10/ghbch2

Eisele, G., Vachon, H., Lafit, G., Tuyaerts, D., Houben, M., Kuppens, P., Myin-Germeys, I., & Viechtbauer, W. (2022). A mixed-method investigation into measurement reactivity to the experience sampling method: The role of sampling protocol and individual characteristics. *Psychological Assessment*, *35*, 68–81. https://doi.org/10.1037/pas0001177

Entwistle, C., Horn, A. B., Meier, T., Hoemann, K., Miano, A., & Boyd, R. L. (2023). Natural emotion vocabularies and borderline personality disorder. *Journal of Affective Disorders Reports*, *14*, 100647. https://doi.org/10.1016/j.jadr.2023.100647

Erbas, Y., Ceulemans, E., Kalokerinos, E. K., Houben, M., Koval, P., Pe, M. L., & Kuppens, P. (2018). Why I don't always know what I'm feeling: The role of stress in within-person fluctuations in emotion differentiation. *Journal of Personality and Social Psychology*, *115*(2), 179–191. https://doi.org/10.1037/pspa0000126

Erbas, Y., Kalokerinos, E., Kuppens, P., van Halem, S., & Ceulemans, E. (2021). Momentary emotion differentiation: The derivation and validation of a framework to study within-person fluctuations in emotion differentiation. *Assessment*, *29*, 700–716. https://doi.org/10.1177/1073191121990089

Faraji-Rad, A., Tamaddoni, A., & Jebeli, A. (2024). *Coping through precise labeling of emotions: A deep learning approach to studying emotional granularity in consumer reviews*. https://doi.org/10.31234/osf.io/hjtfn

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Feldman, L. A. (1993). Distinguishing depression and anxiety in self-report: Evidence from confirmatory factor analysis on nonclinical and clinical samples. *Journal of Consulting and Clinical Psychology*, *61*(4), 631–638.

Grossmann, I., Oakes, H., & Santos, H. C. (2019). Wise reasoning benefits from emodiversity, irrespective of emotional intensity. *Journal of Experimental Psychology: General*, *148*(5), 805–823. https://doi.org/10.1037/xge0000543

Hansen, C. H., & Shantz, C. A. (1995). Emotion-specific priming: Congruence effects on affect and recognition across negative emotions. *Personality and Social Psychology Bulletin*, *21*(6), 548–557. https://doi.org/10.1177/0146167295216001

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, *40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Hegefeld, H. M., Satpute, A. B., Ochsner, K. N., Davidow, J. Y., & Nook, E. C. (2023). Fluency generating emotion words correlates with verbal measures but not emotion regulation, alexithymia, or depressive symptoms. *Emotion*, *23*(8), 2259–2269. https://doi.org/10.1037/emo0001229

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10/c9j35b

Hoemann, K., Barrett, L. F., & Quigley, K. S. (2021). Emotional granularity increases over the course of experience sampling: Methodological and individual factors influence how much. *Frontiers in Psychology*, *12*, Article 704125. https://doi.org/10.3389/fpsyg.2021.704125

Hoemann, K., Gendron, M., Crittenden, A. N., Mangola, S. M., Endeko, E. S., Dussault, È., Barrett, L. F., & Mesquita, B. (2023). What we can learn about emotion by talking with the Hadza. *Perspectives on Psychological Science*, *19*, 173–200. https://doi.org/10.1177/17456916231178555

Hoemann, K., Khan, Z., Feldman, M. J., Nielson, C., Devlin, M., Dy, J., Barrett, L. F., Wormwood, J. B., & Quigley, K. S. (2020). Context-aware experience sampling reveals the scale of variation in affective experience. *Scientific Reports*, *10*, 12459. https://doi.org/10.1038/s41598-020-69180-y

Hoemann, K., Khan, Z., Kamona, N., Dy, J., Barrett, L. F., & Quigley, K. S. (2021). Investigating the relationship between emotional granularity and cardiorespiratory physiological activity in daily life. *Psychophysiology*, *58*(6), Article e13818. https://doi.org/10.1111/psyp.13818

Hoemann, K., Lee, Y., Kuppens, P., Gendron, M., & Boyd, R. L. (2023). Emotional granularity is associated with daily experiential diversity. *Affective Science*, *4*,291–306. https://doi.org/10.1007/s42761-023-00185-2

Hoemann, K., Vicaria, I. M., Gendron, M., & Stanley, J. T. (2021). Introducing a face sort paradigm to evaluate age differences in emotion perception. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *76*(7), 1272–1281. https://doi.org/10.1093/geronb/gbaa038

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*(6), 1744–1756. https://doi.org/10.3758/s13423-016-1053-2

Ip, K. I., Yu, K., & Gendron, M. (2023). Emotion granularity, regulation, and their implications in health: Broadening the scope from a cultural and developmental perspective. *Emotion Review*. Advance online publication. https://doi.org/10.1177/17540739231214564

Johnson-Laird, P. N., & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition and Emotion*, *3*(2), 81–123. https://doi.org/10.1080/02699938908408075

Kalokerinos, E. K., Erbas, Y., Ceulemans, E., & Kuppens, P. (2019). Differentiate to regulate: Low negative emotion differentiation is associated with ineffective use but not selection of emotion-regulation strategies. *Psychological Science*, *30*(6), 863–879. https://doi.org/10.1177/0956797619838763

Kang, S.-M., & Shaver, P. R. (2004). Individual differences in emotional complexity: Their psychological implications. *Journal of Personality*, *72*(4), 687–726. https://doi.org/10.1111/j.0022-3506.2004.00277.x

Kashdan, T. B., Barrett, L. F., & McKnight, P. E. (2015). Unpacking emotion differentiation: Transforming unpleasant experience by perceiving distinctions in negativity. *Current Directions in Psychological Science*, *24*(1), 10–16. https://doi.org/10.1177/0963721414550708

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173.

Kross, E., Verduyn, P., Boyer, M., Drake, B., Gainsburg, I., Vickers, B., Ybarra, O., & Jonides, J. (2019). Does counting emotion words on online social networks provide a window into people's subjective experience of emotion? A case study on Facebook. *Emotion*, *19*(1), 97–107. https://doi.org/10.1037/emo0000416

Kuppens, P., Dejonckheere, E., Kalokerinos, E. K., & Koval, P. (2022). *Some recommendations on the use of daily life methods in affective science* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/y4aqh

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*(6), 1042–1060. https://doi.org/10.1037/a0020962

Labouvie-Vief, G., & Medler, M. (2002). Affect optimization and affect complexity: Modes and styles of regulation in adulthood. *Psychology and Aging*, *17*(4), 571.

Lane, R. D., Quinlan, D. M., Schwartz, G. E., Walker, P. A., & Zeitlin, S. B. (1990). The Levels of Emotional Awareness Scale: A cognitive-developmental measure of emotion. *Journal of Personality Assessment*, *55*(1–2), 124–134. https://doi.org/10.1080/00223891.1990.9674052

Larsen, J. T., & McGraw, A. P. (2011). Further evidence for mixed emotions. *Journal of Personality and Social Psychology*, *100*(6), 1095–1110. https://doi.org/10.1037/a0021846

Larsen, R. J. (1987). The stability of mood variability: A spectral analytic approach to daily mood assessments. *Journal of Personality and Social Psychology*, *52*(6), 1195–1204.

Larsen, R. J., & Diener, E. (1987). Affect intensity as an individual difference characteristic: A review. *Journal of Research in Personality*, *21*, 1–39.

Li, Y., Masitah, A., & Hills, T. T. (2020). The Emotional Recall Task: Juxtaposing recall and recognition-based affect scales. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(9), 1782–1794. https://doi.org/10/ggtmhq

Li, Z., Lu, H., Liu, D., Yu, A. N. C., & Gendron, M. (2023). *Emotional event perception is related to lexical complexity and emotion knowledge*. PsyArXiv. https://doi.org/10.31234/osf.io/vfsc4

MacVittie, A., Kochanowska, E., Kam, J., Allen, L., Mills, C., & Wormwood, J. B. (under review). *First-person thought is associated with body awareness in daily life: Evidence from ecological momentary assessment studies*.

MacVittie, A., Kochanowska, E., Kam, J., Allen, L., Mills, C., & Wormwood, J. B. (in press). *Momentary awareness of body sensations is associated with concurrent affective experience*. *Emotion*.

MATLAB. (2018). *9.5.0.1033004 (R2018b Update 2)* [Computer software]. The Mathworks.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *Manual: Profile of mood states*. Educational and Industrial Testing Services.

Mehta, P., & Pandya, . (2020). A review on sentiment analysis methodologies, practices and applications. *International*

Journal of Scientific and Technology Research, *9*(2), 601–609.

Mestdagh, M., Pe, M., Pestman, W., Verdonck, S., Kuppens, P., & Tuerlinckx, F. (2018). Sidelining the mean: The relative variability index as a generic mean-corrected variability measure for bounded variables. *Psychological Methods*, *23*(4), 690–707. https://doi.org/10.1037/met0000153

Mestdagh, M., Verdonck, S., Piot, M., Niemeijer, K., Kilani, G., Tuerlinckx, F., Kuppens, P., & Dejonckheere, E. (2023). m-Path: An easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioral research and clinical practice. *Frontiers in Digital Health*, *5*, Article 1182175. https://doi.org/10.3389/fdgth.2023.1182175

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465. https://doi.org/10/f45xmb

Nowlis, V. (1965). Research with the Mood Adjective Check List. In S. S. Tompkins & C. E. Izard (Eds.), *Affect, cognition, and personality: Empirical studies* (pp. 352–389). Springer.

Nowlis, V., & Nowlis, H. H. (1956). The description and analysis of mood. *Annals of the New York Academy of Sciences*, *65*(4), 345–355. https://doi.org/10.1111/j.1749-6632.1956.tb49644.x

Olthof, M., Hasselman, F., Aas, B., Lamoth, D., Scholz, S., Daniels-Wredenhagen, N., Goldbeck, F., Weinans, E., Strunk, G., Schiepek, G., Bosman, A. M. T., & Lichtwarck-Aschoff, A. (2023). The best of both worlds? General principles of psychopathology in personalized assessment. *Journal of Psychopathology and Clinical Science*, *132*(7), 808–819. https://doi.org/10.1037/abn0000858

Ortony, A., Clore, G. L., & Foss, M. A. (1987). The referential structure of the affective lexicon. *Cognitive Science*, *11*(3), 341–364. https://doi.org/10.1207/s15516709cog1103_4

O'Toole, M. S., Renna Megan, E., Elkjær, E., Mikkelsen, M. B., & Mennin, D. S. (2020). A systematic review and meta-analysis of the association between complexity of emotion experience and behavioral adaptation. *Emotion Review*, *12*(1), 23–38. https://doi.org/10/ggcb24

Ottenstein, C., & Lischetzke, T. (2019). Development of a novel method of emotion differentiation that uses open-ended descriptions of momentary affective states. *Assessment*, *27*, 1928–1945. https://doi.org/10.1177/1073191119839138

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., Kuppens, P., & Gotlib, I. H. (2015). Emotion-network density in major depressive disorder. *Clinical Psychological Science*, *3*(2), 292–300. https://doi.org/10.1177/2167702614540645

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS): Depression, anxiety, and anger. *Assessment*, *18*(3), 263–283. https://doi.org/10.1177/1073191111411667

Piot, M., Mestdagh, M., Riese, H., Weermeijer, J., Brouwer, J. M. A., Kuppens, P., Dejonckheere, E., & Bos, F. M. (2022). Practitioner and researcher perspectives on the utility of ecological momentary assessment in mental health care: A survey study. *Internet Interventions*, *30*, 100575. https://doi.org/10.1016/j.invent.2022.100575

Pond, R. S., Jr Kashdan, T. B., DeWall, C. N., Savostyanova, A., Lambert, N. M., & Fincham, F. D. (2012). Emotion differentiation moderates aggressive tendencies in angry people: A daily diary analysis. *Emotion*, *12*(2), 326–337. https://doi.org/10.1037/a0025762

Quoidbach, J., Berry, E. V., Hansenne, M., & Mikolajczak, M. (2010). Positive emotion regulation and well-being: Comparing the impact of eight savoring and dampening strategies. *Personality and Individual Differences*, *49*(5), 368–373. https://doi.org/10.1016/j.paid.2010.03.048

Quoidbach, J., Gruber, J., Mikolajczak, M., Kogan, A., Kotsou, I., & Norton, M. I. (2014). Emodiversity and the emotional ecosystem. *Journal of Experimental Psychology: General*, *143*(6), 2057–2066. https://doi.org/10.1037/a0038025

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *Quarterly Journal of Experimental Psychology*, *68*(8), 1584–1598. https://doi.org/10.1080/17470218.2014.941296

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using Siamese bert-networks. *arXiv Preprint arXiv:1908.10084*.

Reimers, N., & Gurevych, I. (2020). *Making monolingual sentence embeddings multilingual using knowledge distillation*.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178.

Salarian, A. (2016). *Intraclass Correlation Coefficient (ICC)* (1.3.0.0) [MATLAB]. https://www.mathworks.com/matlabcentral/fileexchange/22099-intraclass-correlation-coefficient-icc

Seah, T. H. S., & Coifman, K. (2021). Emotion differentiation and behavioral dysregulation in clinical and non-clinical samples: A meta-analysis. *Emotion*, *22*, 1686–1697. https://doi.org/10.1037/emo0000968

Snippe, E., Smit, A. C., Kuppens, P., Burger, H., & Ceulemans, E. (2023). Recurrence of depression can be foreseen by monitoring mental states with statistical process control. *Journal of Psychopathology and Clinical Science*, *132*, 145–155. https://doi.org/10.1037/abn0000812

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, *33*, 16857–16867.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.

Starr, L. R., Hershenberg, R., Li, Y. I., & Shaw, Z. A. (2017). When feelings lack precision: Low positive and negative emotion differentiation and depressive symptoms in daily life. *Clinical Psychological Science*, *5*(4), 613–631. https://doi.org/10.1177/2167702617694657

Stone, A. A. (1981). The association between perceptions of daily experiences and self-and spouse-rated mood. *Journal of Research in Personality*, *15*(4), 510–522. https://doi.org/10.1016/0092-6566(81)90047-7

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine*, *16*(3), 199–202.

Sun, J., Kern, M. L., Schwartz, H. A., Son, Y., & Vazire, S. (2019). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, *118*(2), 364–387.

Suvak, M. K., Litz, B. T., Sloan, D. M., Zanarini, M. C., Barrett, L. F., & Hofmann, S. G. (2011). Emotional granularity and borderline personality disorder. *Journal of Abnormal Psychology*, *120*(2), 414–426. https://doi.org/10.1037/a0021808

Thompson, R. J., Liu, D. Y., Sudit, E., & Boden, M. (2021). Emotion differentiation in current and remitted major depressive disorder. *Frontiers in Psychology*, *12*, Article 685851. https://doi.org/10.3389/fpsyg.2021.685851

Tomko, R. L., Lane, S. P., Pronove, L. M., Treloar, H. R., Brown, W. C., Solhan, M. B., Wood, P. K., & Trull, T. J. (2015). Undifferentiated negative affect and impulsivity in borderline personality and depressive disorders: A momentary perspective. *Journal of Abnormal Psychology*, *124*(3), 740–753. https://doi.org/10.1037/abn0000064

Tov, W., Ng, K. L., Lin, H., & Qiu, L. (2013). Detecting well-being via computerized content analysis of brief diary entries. *Psychological Assessment*, *25*(4), 1069–1078. https://doi.org/10.1037/a0033007

Tugade, M. M., Fredrickson, B. L., & Barrett, L. F. (2004). Psychological resilience and positive emotional granularity: Examining the benefits of positive emotions on coping and health. *Journal of Personality*, *72*(6), 1161–1190. https://doi.org/10.1111/j.1467-6494.2004.00294.x

van de Leemput, I. A., Wichers, M., Cramer, A. O., Borsboom, D., Tuerlinckx, F., Kuppens, P., Van Nes, E. H., Viechtbauer, W., Giltay, E. J., & Aggen, S. H. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, *111*(1), 87–92. https://doi.org/10.1073/pnas.1312114110

Van Rensbergen, B., De Deyne, S., & Storms, G. (2016). Estimating affective word covariates using word association data. *Behavior Research Methods*, *48*(4), 1644–1652. https://doi.org/10/f9pgb9

Vine, V., Boyd, R. L., & Pennebaker, J. W. (2020). Natural emotion vocabularies as windows on distress and well-being. *Nature Communications*, *11*(1), Article 1. https://doi.org/10/ghbwk3

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Watson, D., Anna, L., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.

Watson, D., & Clark, L. A. (1994). *The PANAS-X: Manual for the positive and negative affect schedule-expanded form*. https://doi.org/10.17077/48vt-m4t2

Weermeijer, J.D.M., Wampers, M., De Thurah, L., Bonnier, R., Piot, M., Kuppens, P., Myin-Germeys, I., & Kiekens, G. (2023). *Usability of the experience sampling method in specialized mental health care: Pilot evaluation study. JMIR Formative Research*, *7*(1), e48821

Weidman, A. C., Sun, J., Vazire, S., Quoidbach, J., Ungar, L. H., & Dunn, E. W. (2019). (Not) hearing happiness: Predicting fluctuations in happy mood from acoustic cues using machine learning. *Emotion*, *20*, 642–658.

Widdershoven, R. L., Wichers, M., Kuppens, P., Hartmann, J. A., Menne-Lothmann, C., Simons, C. J., & Bastiaansen, J. A. (2019). Effect of self-monitoring through experience sampling on emotion differentiation in depression. *Journal of Affective Disorders*, *244*, 71–77. https://doi.org/10/gfj6cz

Wilhelm, F. H., & Grossman, P. (2010). Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. *Biological Psychology*, *84*(3), 552–569. https://doi.org/10.1016/j.biopsycho.2010.01.017

Williams, G. E., & Uliaszek, A. A. (2021). Measuring negative emotion differentiation via coded descriptions of emotional experience. *Assessment*, *29*, 1144–1157. https://doi.org/10/gjp3k5

Wingfield, C., & Connell, L. (2023). Sensorimotor distance: A grounded measure of semantic similarity for 800 million concept pairs. *Behavior Research Methods*, *55*(7), 3416–3432. https://doi.org/10.3758/s13428-022-01965-7

Wotschack, C., & Klann-Delius, G. (2013). Alexithymia and the conceptualization of emotions: A study of language use and semantic knowledge. *Journal of Research in Personality*, *47*(5), 514–523. https://doi.org/10.1016/j.jrp.2013.01.011

Wrzus, C., & Neubauer, A. B. (2022). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, *30*(3), 825–846. https://doi.org/10.1177/10731911211067538