This post-print has been peer-reviewed and is accepted for publication in *Psychological Assessment*.

For Better or for Worse? Visualizing Previous Intensity Levels Improves Emotion (Dynamic) Measurement in Experience Sampling

Egon Dejonckheere[1,2]

Ine Penne[1]

Leontien Briels[1]

Merijn Mestdagh[1]

[1] KU Leuven – Faculty of Psychology and Educational Sciences

[2] Tilburg University – Department of Medical and Clinical Psychology

Correspondence concerning this article should be addressed to Egon Dejonckheere, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, Leuven, 3000, Belgium. E-mail: egon.dejonckheere@kuleuven.be. We made all data, materials and code (R and MATLAB) publicly available at the Open Science Framework: https://osf.io/xturh/.

The preparation of this article was supported by the Research Fund of KU Leuven (C14/19/054). ED and MM are post-doctoral research fellows supported by the Fund for Scientific Research, Flanders (FWO; 1256221N, 1210621N). The funding sources had no involvement in the data collection, analyses or results. The authors declare that they have no conflicting interests.

ED and MM played an equal role in the conceptualization of the study, the interpretation of results, funding acquisition, project administration and supervision. IP and LB collected the data for this experiment. ED wrote the original draft. MM took care of the data curation and formal analyses. All authors reviewed and edited the manuscript and approved of this final version.

Abstract

It is a long known reality that humans have difficulty to accurately rate the absolute intensity of internal experiences, yet the predominant way experience sampling (ESM) researchers assess participants' momentary emotion levels is by means of absolute measurement scales. In a daily-life experiment (n = 178), we evaluate the efficacy of two alternative assessment methods that should solicit a simpler, relative emotional evaluation: (1) visualizing a relative anchor point on the absolute rating scale that depicts people's previous emotion rating, and (2) phrasing emotion items in a relative way by asking for a comparison with earlier emotion levels, using a relative rating scale. Determining five quality criteria relevant for ESM, we conclude that a visual 'Last' anchor significantly improves emotion measurement in daily life: (a) Theoretically, this method has the best perceived user experience, as people, for example, find it the easiest and most accurate way to rate their momentary emotions. Methodologically, this type of measurement generates ESM time series that (b) exhibit less measurement error, produce person-level emotion dynamic measures that are (c) often more stable, and in a few cases show stronger (d) univariate and (e) incremental relations with external criteria like neuroticism and borderline personality (e.g., emotional variability). In sum, we see value in the addition of a relative 'Last' anchor to absolute measurement scales of future ESM studies on emotions, as it structures the ambiguous rating space and introduces more standardization within and between individuals. In contrast, using relatively phrased emotion items is not recommended.

Public Significance Statement

When investigating emotional fluctuations in daily life, most researchers repeatedly request absolute emotion ratings, but accurate intensity scores are difficult to provide in the absence of a clear reference point. Here, we visualize people's previous emotion rating on the measurement scale, and show that this relative anchor point improves real-life emotion measurement.

Keywords: Ecological Momentary Assessment, Affect dynamics, Relative emotion ratings, Emotional variability

For Better or for Worse? Visualizing Previous Intensity Levels Improves Emotion (Dynamic) Measurement in Experience Sampling

When interested in the temporal or contextual dynamics of everyday emotion, researchers often use experience sampling methodology (ESM; Larson & Csikszentmihalyi, 1983). In an ESM study on emotions, participants are repeatedly invited to provide an evaluation of their current emotional state as they live their life (e.g., via their smartphone; Houben et al., 2015; Mestdagh & Dejonckheere, 2021; Waugh & Kuppens, 2021). Generally, these emotional assessments take the form of an absolute intensity rating on a bounded measurement scale (e.g., *How happy do you feel right now?* on a 0 [*not at all*] – 100 [*very much*] visual slider; see De Vuyst et al., 2019; Heininga et al., 2019; Kalokerinos et al., 2019; Koval et al., 2013; Loossens et al., 2020; Panaite et al., 2019; Pe et al., 2015; Thompson et al., 2012; Van der Gucht et al., 2019 for some concrete examples). However, absolute and bounded emotion measures have both theoretical and methodological complications, undermining the reliability and validity of the ESM data they generate. Here, we review problems with absolute, bounded rating scales in the context of emotional ESM. In a within-person ESM experiment, we evaluate two proposed alternative methods that could bypass the outlined shortcomings.

Ambiguous Reference Point

First, theoretically, it is recognized that people struggle to provide accurate absolute intensity ratings to internal subjective experiences in the absence of a clear reference point (Miller, 1956; Stewart et al., 2005). In fact, some scholars have argued that each subjective evaluation is inherently relational (e.g., Helson, 1947; Lockhead, 2004), leading humans to heuristically turn the complex task of absolute judgments into an exercise of comparison against some prespecified reference point (Yannakakis et al., 2017). In the case of emotions, it is obvious that continuous rating scales (e.g., visual analog sliders that generate interval data) provide no tangible references to interpret the corresponding numerical values, except for the

qualitative labels at both scale ends. How one should structure the ambiguous space in between these delineating end points is up to the individual interpretation of each participant (e.g., *How much happiness does 42 reflect on a 0 – 100 happiness scale?*; Aitken, 1969), and even within individuals research suggests that this interpretation changes over time (e.g., initial elevation effects; Shrout et al., 2018). To bypass this problem, ESM researchers could rely on Likert scales to evaluate people's momentary emotions, as these scales typically have more concrete labels associated with each rating point (e.g., 1 [*not at all*] – 2 [*a little*] – 3 [*moderately*] – 4 [*quite a bit*] – 5 [*extremely*]; Nadler et al., 2015). Although labelling makes the rating space less ambiguous (Svensson, 2000), the reference problem remains (e.g., What is *a little* happiness on 5-point Likert scale?), and it is unclear whether the limited range of answer options allows for sufficient emotional intensity differentiation (Simms et al., 2019). Even more worrisome is that the ordinal nature of Likert-type intensity data prohibits the use of many conventional (parametric) analysis methods in ESM (e.g., calculating the mean and standard deviation [SD] from an ordinal time series is not warranted; Jamieson, 2004).

When participants in an ESM study on emotions impose structure to absolute rating scales by means of a relative reference point (Yannakakis et al., 2021), it is crucial for fair between-person comparisons that all participants use the same reference type. Similarly, at the within-person level, consistently adopting the same type of reference may positively impact the information value of the resulting emotional time series. There are many reference types that could serve as a basis for emotion comparison, and depending on the category that one selects, people's momentary intensity ratings will likely vary (Kahneman, 1992). For example, when prompted to provide a momentary happiness rating shortly after an argument with a friend, you could derive your current experienced happiness from a comparison with your ideal happiness levels (*How happy do you want to feel in this situation?*; e.g., Tsai, 2017), your self- or other-ought happiness levels (*How happy do you think you should feel in this situation* [*according to*

others]*?*; e.g., Dejonckheere et al., 2017; Thompson et al., 2016), the happiness levels of other people (*How happy do others feel in this situation?*; Yzerbyt et al., 2006), your average level of happiness (*How happy do you usually feel?*; e.g., Kuppens, Oravecz, et al., 2010), your previous happiness intensity (*How happy were you before this argument?*; e.g., Russell & Lanius, 1984), and so on.

To aid ESM participants in the rating of their momentary emotions, researchers could mark the momentary value of a certain reference category on the absolute rating scale. However, in many instances this reference information is not readily available. For example, the idiosyncratic aspect of ESM participants' personal ecologies makes it impossible to learn how others would feel in the exact same rating situation. Other momentary reference values, such as ideal or ought momentary emotions, only reside in people's mind and require a separate assessment themselves (Sims et al., 2015). In contrast, how participants previously felt, is easily retrieved due to the repetitive nature of the momentary ESM surveys.

Using people's previous emotional state as a basis for comparison in absolute emotion intensity ratings nicely aligns with the central proposition of *adaptation level theory* (Helson, 1947), which posits that stimulus judgements are predominantly determined by the (weighted mean of) stimuli we encountered earlier. Together, these stimuli make up a changing adaptation level that serves as a basis for departure for any human experience (Helson, 1964). Applied to emotional judgements, Russell and Lanius (1984) demonstrated experimentally how previously presented affect-laden images serve as an important source of information for subsequent affective evaluations. However, to what extent the presentation of an emotional adaptation level (in the form of an anchor that depicts one's previous emotional state) facilitates absolute emotion ratings in the context of everyday life, currently remains uninvestigated.

Bound-related Problems

A second line of (methodological) complications with absolute emotion ratings in ESM stems from the boundedness of the measurement scales (Mestdagh et al., 2018). The scale end points naturally constrain the rating space that can be used to evaluate one's momentary emotion level, but the scale's absolute quality also expects participants to have a decent notion of all the forthcoming rating situations in order to correctly calibrate the scale. To illustrate, the previous argument with a friend may have lead you to select the maximum score on a 0 – 100 sadness scale. However, when shortly thereafter you learn that your friend tragically died in a car accident, you are in no position to provide a more extreme sadness rating that accurately reflects your current emotional state. Ceiling and floor effects as a result of bounded absolute emotion scales undermine the authentic rating of beyond-scale experiences (Wang et al., 2008). This impedes a correct representation of people's actual within-person emotional fluctuations.

Another issue related to floor and ceiling effects in bounded emotion ratings is the fact that these scale attenuations introduce a structural relation between a person's mean level of emotion and the maximum amount of within-person emotional variability one can display (Baird et al., 2006; Eid & Diener, 1999). When people consistently rate the absolute intensity of their momentary emotions as either very high or very low, their mean levels will be situated close to one of the scale boundaries. Due to a constraint of the scale's floor or ceiling, these individuals cannot show as much variability in their absolute emotion ratings as compared to people who have a mean emotion level around the scale midpoint (where higher amounts of emotional variability are possible; Mestdagh et al., 2018; Vanbelle & Lesaffre, 2018). The collinearity between people's emotional mean and variability is particularly strong for negative emotions, as for many people the day-to-day intensity of these negative states is usually low, producing a strongly (right-)skewed distribution (Trampe et al., 2015). As a consequence, assessing infrequent (negative) emotional states with bounded absolute intensity scales may lead to spurious associations between emotional variability and other person-level variables when this critical mean-variability dependency is not taken into account. Such a confound has been established in the domain of personality with respect to neuroticism (e.g., Hisler et al., 2020; Kalokerinos et al., 2020; Wendt et al., 2020), and in mental well-being research regarding borderline symptomatology (e.g., Dejonckheere et al., 2019; Houben et al., 2021). Theoretically, both of these trait criteria are thought to

be defined by strong emotional ups and downs (Carpenter & Trull, 2013; Eysenck, 1985). However, empirically, a mean-confound leaves researchers in the dark about the true explanatory role of emotional variability in these person-level characteristics (Mestdagh et al., 2018).

To allow for the truthful assessment of beyond-scale experiences and a reduced mean-confound of emotional variability, ESM researchers may turn to an evaluation of people's momentary emotions that is entirely relative in its phrasing. By framing the intensity assessment of emotions in a relative way, participants are invited to compare their current emotion levels to how they previously felt (e.g., *compared to the previous measurement, how happy do you feel right now?* on a scale with -50 [*much less happy*] – 0 [*equally happy*] – +50 [*much happier*]). In this way, participants always have the opportunity to accurately provide momentary emotion ratings that are more extreme than the previous one(s), which could effectively eliminate the floor or ceiling restraints that are imposed by the bounds of absolute measures. Since relative intensity scales bypass the impossible task of having to anticipate all possible rating situations, there is no need for participants to calibrate the rating scale in advance.

Related to the problematic mean-confound in people's within-person emotional variability, it is obvious that relative emotion scales allow for a more direct and natural assessment of people's moment-to-moment emotional fluctuations[1]. Mathematically, to summarize a group of scores, computing the first (statistical) moment of a series of absolute intensity ratings results in a person's average intensity (i.e., one's mean level of emotion). This is followed by the second moment, which represents a person's average deviation from that mean level (i.e., one's variability in emotion; Dejonckheere et al., 2019). However, when adopting relative rating scales, the interpretation of these statistical moments changes: The first moment of a series of (the absolute value of) relative scores now represents a person's average deviation, while the computation of one's average intensity requires a more complex derivation. By framing emotional assessments in a relative way, the idea is that the explanatory power of emotional variability as the first moment may be stronger than people's mean levels of emotion in the prediction

[1] In fact, a comparative assessment of momentary emotions in relation to how people previously felt directly corresponds to their level of emotional instability, which represents the average emotional intensity change between two successive measurement occasions (in an absolute context, this is often operationalized as the root of the mean squared successive difference [RMSSD] between consecutive emotion ratings; Jahng et al., 2008).

of trait variables like neuroticism or borderline symptoms. If so, relative emotion ratings enable a more direct evaluation of emotional variability's predictive validity, less confounded by the predominant predictive power of people's mean levels of emotion.

However, this entire rationale is based on the proposition that participants have a good recollection of their prior emotional state, an idea that has no scientific consensus (Retkoceri, 2022). For instance, while some scholars claim that people may directly retrieve emotional experiences from their memory (e.g., LeDoux, 1996), others contend that the experiential circumstances need to be cognitively reconstructed first (e.g., Robinson & Clore, 2002). Similarly, diverging theoretical positions emerge as to whether previously experienced emotions are stored indelibly (e.g., as seen in the reinstatement of extinguished conditioned emotion after exposure to an event that induces similar feelings; Halladay et al., 2012), or whether recollected emotions are susceptible to forgetting or recall biases (e.g., supported by the observation that people's recalled emotions are partly shaped by their current appraisals of the emotion-eliciting event; Levine et al., 2001). In essence, this debate stems from whether researchers assume that people retrieve previous emotional experiences either from their *implicit* (i.e., non-declarative) or *explicit* (e.g., episodic or semantic) memory (e.g., see Sanna & Chang, 2006 for a comprehensive review of the literature on [mis]remembering emotions).

The Current Study

To examine whether relative emotion assessments could remediate the theoretical and methodological problems associated with absolute, bounded emotion measures in ESM, we conducted a within-person experiment. Besides using the standard absolute intensity scales (ABS), we instructed participants to rate the intensity of their real-life momentary emotions by means of an absolute scale with a relative anchor that indicated the intensity of their previous measurement (ABS-REL), and a rating scale that was entirely relative due to the item's phrasing (REL; see Figure 1). We evaluated the potential of each measurement method on the basis of five quality criteria that are relevant for time series research. Theoretically, we explored participants' subjective user experience with each of the measurement methods (one criterion). Methodologically, we assessed the internal (one criteria) and external validity (three criteria) of each measurement method's emotion time series. This validity distinction in ESM was proposed by Dejonckheere and Mestdagh (2021), and refers to different (within-

versus between-person) predictive properties of a time series. Consequently, the current investigation pertain both levels of analysis.

[Figure 1 around here]

User Experience

Theoretically, it is important that emotion measures in ESM are face valid, not only in terms of item content, but also in terms of assessment procedure (Nevo, 1985). Consequently, how ESM researchers assess the intensity of momentary emotions should align with the intuitive notion that people have about their emotions. If a relative comparison indeed lies at the basis of people's emotional experience (e.g., Kahneman, 1992; Russell & Lanius, 1984), then we may expect that participants prefer relative over absolute intensity rating scales (of course the reverse is not necessarily true). However, an emotional assessment that is entirely relative also requires considerable cognitive effort, as participants need to actively recollect their previous emotional state (Carlsson, 1983). It is possible that the burden of this mental operation undermines the favorable user experience of purely relative rating scales.

Internal Validity of ESM Time Series

Methodologically, a first criterion relates to the ability of emotional assessments to accurately describe the (within-person) ups and downs of people's emotional life. When a series of emotion ratings does not correctly represent the true natural fluctuations in people's emotions, and intraindividual changes in observed emotion ratings are mainly driven by random fluctuations caused by measurement error, the true dynamic qualities of people's emotions are not revealed (Schuurman et al., 2015). Likewise, detecting meaningful within-person associations with other momentary constructs will be harder, because noisy ratings are known to attenuate statistical power at the within-person level (Dejonckheere et al., 2022). Therefore, it is important to determine the internal validity of an emotion time series, which refers to the extent with which we can accurately predict the true future dynamic trajectory of an individual's emotional state based on previous emotional assessments.

An evident way to maximize the internal validity of a series of emotion ratings is to reduce the error variance related to the measurement process (Dejonckheere & Mestdagh, 2021). We will thus investigate whether relative emotion assessment result in more precise measurement precision. In our hypotheses, we differentiate between instances where people's true emotion level steadily persists across

measurement occasions versus meaningfully fluctuates. First, when the intensity of an emotion did not change across assessments, we argue that it is easier to indicate so in a relative framework. Specifically, in an ABS-REL context, participants can simply select the anchor position to indicate that their emotion levels remained constant. In a REL context, this can be accomplished by marking the scale's middle point, additionally assuming that participants correctly remember how they previously felt (which is debatable, Retkoceri, 2022). If true, the presentation of these explicit reference markers on ABS-REL and REL rating scales could result in more accurate ratings for similar true emotion levels compared to ABS scales, where these concrete guiding points are not presented. Second, however, at times when people's emotions do meaningfully change across assessments, REL ratings may introduce more measurement error variance to people's emotion time series than ABS-REL ratings. The rationale for this hypothesis is that when REL ratings are transformed to their absolute counterpart (see *Preprocessing ESM Time Series* paragraph for more information), the measurement error that is inherent to each emotion assessment inevitably accumulates over time, creating observed absolute emotion ratings that increasingly diverge from their true score (i.e., a *drift* process; see Supplementary Materials 1 for some simulated and empirical examples). In ABS-REL ratings, this problematic dependency is (partially) minimized by the absolute qualities of the measurement scale. Taken together, we expect measurement error variance to be the lowest in ABS-REL time series, followed by REL time series, and highest in traditional ABS time series.

External Validity of ESM Time Series

Second, when the measurement error of people's emotion time series is reduced, we may expect that (dynamic) summaries computed from these momentary emotion ratings show stronger meaningful relations with external person-level variables like neuroticism or borderline symptoms (Dejonckheere et al., 2020). The ability to predict independent (between-person) criteria refers to the external validity of a series of emotion ratings (Dejonckheere & Mestdagh, 2021). Various (dynamical) time series aggregates have been linked to individual differences in neuroticism (Hisler et al., 2020; Miller et al., 2009; Suls et al., 1998) and borderline symptomatology (Houben et al., 2015, 2021; Jahng et al., 2008). In addition to people's mean levels of positive and negative emotion, the variability in these affective states (operationalized as the within-person SD; Koval et al., 2013), their inertia (operationalized as the

observed auto-regressive [AR] relation; Kuppens, Allen, et al., 2010), and their instability (operationalized as the root of the mean squared successive difference between consecutive emotion ratings [RMSSD]; Jahng et al., 2008) have received most attention in the emotion dynamic literature, and will be our focus in this study.

Reliability of Emotion (Dynamic) Measures. To enable an effective prediction of external criteria, a first prerequisite is that emotion (dynamic) measures as between-person constructs show sufficient reliability, as this aspect determines the upper bound of their association with other criteria (Wendt et al., 2020). This means that summaries of people's emotion time series should not drastically change when a slightly different sample of their momentary emotion ratings is considered. If emotion (dynamic) aggregates are indeed valid person-level constructs, we can expect sufficient short-term consistency (Mejía et al., 2014). By computing split-half correlations between emotion (dynamic) measures derived from odd and even assessment days, previous research established high reliabilities for people's emotional means and variability using ABS ratings (Wendt et al., 2020). In contrast, the reliability of emotion dynamic measures that convey temporal information (i.e., instability and inertia) was substantially lower. Here, we hypothesize that this decrease in reliability for emotional instability and inertia may be countered when relying on REL and ABS-REL ratings, respectively. Compared to ABS items, the phrasing of REL items provides a direct assessment of emotional instability (which does not require a complex transformation of the original ESM data; see Footnote 1). In ABS-REL ratings, the relative 'Last' anchor may stimulate a visual comparison with earlier emotion levels, more intuitively evaluating to what extent an emotional state lingers over time (i.e., the core of emotional inertia; Kuppens, Allen, et al., 2010). This is absent in traditional ABS ratings.

Individual Predictive Value. Next, we may evaluate to what extent these emotion (dynamic) measures individually predict differences in neuroticism or borderline symptoms. In terms of univariate explanatory power, we hypothesize that emotion dynamics derived from a series of relative emotion ratings will show stronger relations with people's neuroticism and borderline symptoms compared to those computed from traditional absolute emotion ratings. Specifically, because REL ratings provide a more explicit and intuitive assessment of people's levels of emotional variability and instability, we argue that this may permit bigger effect sizes in the prediction of these person-level variables. Similarly,

when ABS-REL ratings improve the estimation of people's emotional inertia as a result of stronger serial dependency, this could strengthen its link with neuroticism.

Added Predictive Value above Mean Levels of Emotion. Finally, previous ESM research based on ABS ratings demonstrated that observed effect sizes of emotion dynamic measures in the prediction of well-being outcomes are relatively small compared to static mean levels of emotion (Dejonckheere et al., 2019), and it remains unclear whether absolute measurement practices in traditional ESM obfuscate meaningful associations, or if their role in people's neuroticism and borderline symptoms is truly limited. In terms of added explanatory power above static mean levels of positive or negative emotion, we hypothesize that the mean-variability dependency observed in ABS time series will be weaker in a REL (but not ABS-REL) context, because entirely relative ratings should avoid the bound problem associated with absolute scales. Consequently, when people's emotional mean and variability are not that strongly intercorrelated, we may expect that emotional variability derived from REL ratings considerably contributes to the prediction of neuroticism and borderline symptomatology, above and beyond static mean levels of emotion.

Materials and Methods

Transparency and Openness

The present research was approved by the Social and Societal Ethics Committee of KU Leuven (*Emotional variability in daily life: Absolute versus relative assessment? A comparative study on the validity of both assessment methods.*; G-2020-2798). To reproduce our analyses, the reader may retrieve our data and code (R and MATLAB) from the Open Science Framework (https://osf.io/xturh/). We report our sample determination procedure, all data exclusions, and all manipulations. The study design and analyses were not pre-registered.

Power Analysis, Prescreening and Participant Sample

First, at the within-person level, we aimed to collect around 30 completed momentary observations (*t*) per measurement condition for each participant. This number of ESM assessments is sufficient to obtain stable estimates for the most important emotion (dynamic) measures (i.e., more measurements do not significantly improve the stability of mean and *SD*; Jaso et al., 2021). However, because some missingness in ESM is inevitable, we slightly raised the number of scheduled assessments

per measurement condition (t = 40), anticipating that participants would complete around 75% of the momentary notifications (Vachon et al., 2019). This goal was achieved, as our final sample had an average compliance of 89.85% (*SD* = 8.20%), 90.22% (*SD* = 7.63%) and 89.61% (*SD* = 8.14%) in the ABS, ABS-REL and REL condition, respectively.

Second, to select our sample size at the between-person level, we performed an a priori power analysis. Based on the meta-analytic effect sizes reported in Kalokerinos et al. (2020) and Wendt et al. (2020), we anticipated generally small associations ($R^2 \approx .05$) between the emotion dynamic summaries we computed from people's ESM time series and their person-level neuroticism and borderline scores reported at baseline, when taking into account the explanatory power of mean levels of emotion. To ensure a power of 80% (with $\alpha = .05$), an inclusion of at least 200 participants allowed us to obtain sufficiently precise (and therefore adequately comparable) point estimates of the predictive power of different measurement-based emotion dynamic measures (controlled for mean levels). Thus, aiming to keep the standard error around our point estimates sufficiently small allowed us to observe significant differences between conditions. Evidently, with this expected number of participants, the statistical power to assess the *individual* predictive power of each emotion (dynamic) measures was even higher.

To further improve estimator precision, we adopted a stratified sampling approach to increase the observed between-person heterogeneity in neuroticism profiles (Parsons, 2017). To this end, we first implemented a prescreening phase via Qualtrics, in which interested adults could complete a Dutch version of the Big Five Inventory Neuroticism subscale (BFI-N; Denissen et al., 2008). From an initial pool of 983 eligible individuals (84.1% female; age *M* = 23.59, *SD* = 7.78) we then oversampled people with very low and high neuroticism scores, given that these profiles are typically underrepresented in the general population due to the Gaussian distribution of this trait (Haslam, 2017). Specifically, we divided the observed BFI-N prescreening range (*M* = 3.25, *SD* = 0.72, range = 1.13–4.75 on a 5-point Likert scale) into four approximately equal strata, and randomly contacted at least 50 people within each stratum to take part in our actual experiment. The number of invited participants was slightly higher than our planned sample size to account for attrition during the ESM protocol (Vachon et al., 2019).

In the end, we enrolled 209 people in our study, but excluded 31 participants from our ESM analyses, either due to poor compliance rates (i.e., less than 50% completed measurement prompts in

one or more measurement conditions or less than 75% overall[2]; n = 10) and/or a misunderstanding of the experiment instructions (n = 23)[3]. This left us with a final sample of 178 participants with valid ESM data (81.89% female; age M = 23.05, SD = 7.75). Our study sample comprised both general community members and students. They all provided informed consent.

Finally, we additionally conducted a post-hoc power analysis to evaluate what the minimum obtained statistical power was for all tested within-person differences (e.g., paired samples t-tests in the context of our internal validity evaluations). Here, the smallest significant effect size was a Cohen's d of -0.19, which corresponded with a statistical power of 0.71. For all other between-condition differences we had a statistical power of 0.98 or greater.

Procedure

We collected the data for our ESM experiment in April and May, 2021. Due to the ongoing COVID-19 restrictions, all researcher-participant interactions took place via video calls. After the prescreening phase, we invited eligible participants for an online personal introduction session, in which the general study goal and procedure were explained. In order not to reveal our specific research questions or hypotheses, we broadly framed the study as an effort to capture emotional fluctuations in everyday life. Participants then installed our ESM application m-Path (Mestdagh et al., 2023) on their personal smartphone and provided consent. The introduction session was concluded with the completion of our baseline survey via Qualtrics.

Starting the next day, participants took part in the 12-day ESM protocol. Here, our mobile m-Path app repeatedly prompted participants during their daily routines to rate the intensity of their current emotions, and every four days they had to use a different type of rating scale to provide their responses. Once the ESM phase was finished, participants completed a custom-made User Experience (UX)

[2] Abandoning this (arbitrary) >75% overall compliance criterium did not alter our conclusions.

[3] In the REL condition, several participants simply selected the extremes of the measurement scale, leading us to believe they did not understand that they could indicate more subtle deviations from their previous emotion level. Due to the (listwise) deletion of these cases, the established performance of our REL-based measurement method is likely inflated. However, we deem this not problematic as even these more optimistic REL results are still consistently worse than ABS or ABS-REL. At the same time, these misunderstandings highlight the poor user experience of this method.

questionnaire via Qualtrics that assessed their personal experience with each measurement method.

Finally, in a debriefing session, we explained the actual study aim and informed participants about the reimbursement they would receive based on their personal ESM compliance (financially or in research credits). Participants could earn up to €40 or four credits when they completed at least 75% of the ESM assessments. Each decrease of five or 15% resulted in a deduction of €4 or one credit, respectively. For the final ESM sample, the average compliance rate was 89.93% (*SD* = 6.11%). There were no significant differences in terms of compliance between measurement conditions, *t*'s(177) ≤ 0.90, *p*'s ≥ .37.

Materials

Person-level Surveys

Neuroticism. In the prescreening phase, we assessed participants' neuroticism levels via the BFI Neuroticism subscale (Denissen et al., 2008; John & Srivastava, 1999). Besides the other four principal personality dimensions (i.e., extraversion, agreeableness, conscientiousness and openness to experience), the BFI measures people's neurotic tendency, which refers to a predisposition to experience negative emotions (e.g., *I see myself as someone who can be tense.*), an elevated reactivity to environmental stressors (e.g., *I see myself as someone who remains calm in tense situations.* [reversed]), and a turbulent emotional life (*e.g., I see myself as someone who is emotionally stable, not easily upset.* [reversed]; Tackett & Lahey, 2017). Participants rated eight statements on a 5-point scale ranging from 1 (*Strongly disagree*) to 5 (*Strongly agree*). After rescoring all reversed items, we computed the average item rating to obtain a Neuroticism score, with higher BFI-N scores indicating higher levels of neuroticism (*M* = 2.65; *SD* = 0.79). For two participants, we could not link their neuroticism score retrieved during prescreening to their ESM data, because they had changed their pseudonymizing participation code in the ESM app after study completion (e.g., because they took part in another ESM study). This led to a slight decrease in sample size for the analyses that involve this variable (*n* = 176). The internal consistency of this subscale was .84.

Borderline Symptomatology. During the baseline session, we evaluated participants' self-reported borderline symptoms using the Dutch-translated Personality Assessment Inventory Borderline scale (PAI-Bor; Distel et al., 2009). This 24-item survey evaluates the defining core features of

borderline personality disorder with four subscales: Identity problems (e.g., *I often wonder what I should do with my life.*), relational problems (e.g., *My relationships have been stormy.*), emotional instability (e.g., *My mood can shift quite suddenly*), and self-harm behavior (e.g., *When I'm upset, I typically do something to hurt myself.*; Morey, 1991). Each item was rated on a 4-point scale from 1 (*False, not at all true*) to 4 (*Very true*). Due to ethical constraints we did not assess self-harm. After recoding contra-indicative items, we created a total PAI-Bor score by averaging all item ratings, with higher scores indicating higher levels of self-reported borderline symptomatology (M = 2.30, SD = 0.51). For seven participants, we could not match their PAI-Bor score collected during baseline with their ESM data, because they entered an incorrect pseudonymizing participation code for one of the study parts, and/or changed their code in the ESM app after they were finished with the study. Consequently, this resulted in a smaller sample size for the analyses that involve this variable (n = 171). The total scale's internal consistency was .85.

Experience Sampling Protocol

During the ESM protocol, participants were asked to rate the intensity levels of ten momentary emotions, ten times per day for 12 consecutive days (i.e., 120 ESM surveys in total). We adopted a time-contingent semi-random sampling scheme to prompt participants randomly within ten equal blocks between 9 AM and 9 PM. This approach allowed us to cover an entire day representatively without the induction of measurement anticipation seen in fixed sampling designs (which undermines the ecological validity of ESM data; Dejonckheere & Erbas, 2021). The average time interval between two consecutive prompts was 72 minutes (SD = 30 minutes). When participants did not interact with the momentary prompt within a 5-minute window, an automatic reminder was sent. After 15 minutes, the notification expired and participants could no longer complete that survey.

The ten discrete emotions that participants had to rate at each measurement occasion were *happy, excited, enthusiastic, pleased, relaxed, sad, down, irritated, anxious* and *stressed* (the specific order was randomized across momentary assessments). The selection of these emotion items was based on their natural occurrence in daily life (ensuring sufficient within-person variability; Eisele et al., 2021), and their balanced combination of different valence and arousal levels (the two most prominent dimensions that underlie our emotional experience according to the *circumplex model of affect*; Russell,

1980). In our analyses, we considered global momentary positive [PA] and negative affect [NA] composite scores by averaging all same-valenced emotion items at each measurement occasion. No other items were assessed, and participants could not skip questions. They always had to indicate their response on the slider scale before they could proceed to the next item.

Critically, using a within-person experimental design, we divided the entire ESM phase into three equal block of four days (i.e., 40 ESM surveys per condition), that each used a different measurement scale to record participants' momentary ratings of the emotion items described above. We counterbalanced the sequence of the different measurement conditions across participants to eliminate potential learning or fatigue effects. Unique instructions preceded each measurement condition to highlight the use of a different assessment procedure, and participants explicitly indicated that they understood what was expected of them (but see Footnote 3). In the next paragraphs, we explain how we assessed the same set of emotion items in each measurement condition (see Figure 1 for a visual representation of each rating interface).

Absolute Measurement Condition (ABS). In the ABS condition, we phrased all momentary items in an absolute way (e.g., *How happy do you feel right now?*) and invited participants to rate their current emotional intensity on an absolute measurement scale from 0 (*Not happy at all*) to 100 (*Very happy*; see Panel A in Figure 1). The scale midpoint (50) was the slider pointer's initial position. This type of measurement is used most often in conventional ESM research.

Absolute Measurement Condition with a Relative 'Last' Anchor Point (ABS-REL). In the ABS-REL condition, we phrased all momentary items in an absolute way (e.g., *How happy do you feel right now?*), but stimulated a relative comparison in participants' rating process. This was realized by adding a 'Last' anchor to the absolute measurement scale (ranging from 0 [*Not happy at all*] to 100 [*Very happy*]) that indicated the intensity of participants' previous emotional assessment (see Panel B in Figure 1), a feature we developed specifically for this study. The answer pointer always started at the 'Last' anchor. To avoid emotion ratings as a result of overnight comparisons (which could be distorted by memory biases; Bringmann et al., 2013), we ensured that the first prompt of each day in this condition always presented an ABS rating scale (i.e., without a 'Last' anchor). Furthermore, if participants missed one of the following assessment occasions, the anchor always referred to the last completed assessment.

Missingness did thus not affect anchor visualizations. The anchor merely depicted the intensity of the previous rating, not the timing of the assessment.

Relative Measurement Condition (REL). Finally, in the REL condition, we phrased all momentary items in a relative way (e.g., *Compared to the previous measurement, how happy do you feel right now?*). Consequently, participants compared the intensity of their current emotional state to the (recalled) level of that emotion at the previous assessment (see Panel C in Figure 1), using a relative scale that ranged from -50 (*much less happy*) to 0 (the scale's midpoint; *equally happy*) to +50 (*much happier*). The initial position of the answer pointer corresponded to the scale's midpoint (0). Analogous to the ABS-REL condition, we imposed an ABS rating scale for the first assessment of each day to eliminate overnight comparisons that had unusually long time windows (Bringmann et al., 2013)[4].

User Experience

Finally, we created a study-specific UX instrument to quantify participants' personal preference regarding the different measurement methods they used to rate their momentary emotion levels during the ESM protocol. Specifically, we invited participants to rank the three measurement methods from *best* (1) to *worst* (3) for six criteria: perceived *naturalness*, *ease*, *reliability*, *accuracy*, *quickness* and *re-use preference* (see Supplementary Materials 3 for actual item wording). An average ordinal reliability coefficient of .84 across conditions indicated high consistency among rankings, justifying the interpretation of a homogeneous composite score. In addition, we examined the ranking for each criterium separately, classifying the measurement method that was most frequently ranked first among participants as the best fitting method for that criterium. Finally, to obtain a more objective indication of the *ease* / *quickness* with which participants rated the intensity of their momentary emotions, we also

[4] Compared to ABS, it is possible that the accuracy of REL ratings (and to a smaller extent, ABS-REL ratings) is more affected by occasional missingness, because skipped assessments introduce longer time windows between one's recalled and present emotion levels. Re-estimating these time series' measurement error variance using only consecutive assessments that had a time window shorter than three hours (thus eliminating emotional comparisons that exceeded a 3-hour window), we found that missingness had little effect on the REL and ABS-REL method's performance in the current study (see Supplementary Materials 2). However, in the current study, missingness was generally not problematic. When missingness is considerable, the estimated accuracy of REL (and ABS-REL) ratings may be more undermined nonetheless.

computed their average response time following each measurement method, and compared these between conditions using paired samples t-tests (see Supplementary Materials 4 for these results).

Statistical Analyses

All significance tests were two-sided with an α of .05. To maximize statistical power, we always used all available data for a particular analysis. This resulted in slightly diverging sample sizes for different statistical tests (ESM: n = 178, BFI-N: n = 176, PAI-Bor: n = 171, UX: n = 209).

Preprocessing ESM Time Series

To enable an effective comparison between the emotional means of the different measurement conditions, we first transformed people's REL emotion ratings to their absolute counterpart. To this end, we considered the first absolute intensity rating of each day in this REL condition, and used it as a benchmark to compute absolute emotion scores. Thus, participants' first absolute emotion rating (situated between 0 and 100) allowed us to calculate how subsequent relative emotion ratings (ranging from -50 to +50) resulted in consecutive deviations from this original absolute intensity score (see Supplementary Materials 5 for a simulated example). The theoretical minimum and maximum for these transformed REL ratings was -450 (0 - 9 * -50) and 550 (100 + 9 * 50), respectively, and participants, on average, reported beyond-scale ratings in 28.46% of the REL measurement occasions (i.e., intensity scores that fell outside the absolute 0 – 100 interval). As theoretically expected, the RMSSDs computed from people's transformed REL time series almost correlated perfectly with the mean of (the absolute values of) people's raw REL time series (r PA = 0.99, r NA = 0.99; p's < .001), indicating that these metrics are identical measurement operationalizations of emotional instability (see also Footnote 1). Finally, for all three conditions, we created a PA and NA time series for each participant by averaging all same-valenced emotion ratings at each momentary assessment.

User Experience

For both the total UX score and each UX criterium individually, we evaluated the number of participants that ranked a particular condition as the best fitting measurement method. We used z-tests to determine whether that proportion significantly differed from the proportion of the second most selected measurement condition.

Internal Validity

To evaluate the internal validity of the PA and NA time series in each measurement condition, we estimated the measurement error variance associated with people's momentary emotion ratings. To do so, we assumed that people's observed emotional fluctuations were shaped by a latent auto-regressive model of order 1 (i.e., an AR[1] model) with a white noise term (Schuurman & Hamaker, 2019):

$$Happy_t = a\widetilde{Happy}_{t-1} + \varepsilon_t + \omega_t \quad (1)$$

Conceptualizing the true emotion generating model as a latent AR(1) model is common in the emotion dynamics literature, because its AR parameter *a* is intuitively interpreted as *emotional inertia* (e.g., Koval et al., 2013; Kuppens et al., 2010; Suls et al., 1998). This coefficient conveys to what extent a person's true happiness score ($\widetilde{Happy}$) at time point *t*, for example, can be (linearly) predicted from his or her previous true happiness score at time point *t*-1. It therefore indicates to what extent a person's true emotions are self-predictive and resistant the external influences. In dynamical system terms, this coefficient is reversely related to individual differences in *attractor strength*, the speed with which internal regulatory processes pull an emotion back to its homebase (Chow et al., 2005). Consequently, when external influences do impact a person's true emotions at time point *t*, this cannot be directly modelled through the AR relation, but rather is reflected in the residual ε_t (Schuurman et al., 2015). This error term is called *dynamical* or *innovation* noise because it introduces new information to the system at each *t*, which resonates indirectly through a person's subsequent emotional states via his or her personal AR relation (Dejonckheere & Mestdagh, 2021). Finally, the remaining error variance ω_t then reflects random imperfections associated with the momentary measurement process itself (e.g., fatigue, inattention, etc.). These white noise distortions are restricted to each specific momentary emotion assessment, and their effect does not carry over through the next ratings (separating them from innovation noise; Schuurman et al., 2015).

We estimated measurement error variances for both affective states in each measurement condition, and evaluated the respective sample distributions. We used paired samples t-tests to determine whether mean scores differed significantly across conditions.

External Validity

In each measurement condition, we computed four emotion (dynamic) measures for both people's PA and NA time series. First, we considered participants' mean levels of emotion,

operationalized as the average intensity rating across all completed momentary assessments for a particular condition (M; Larson et al., 1980). Second, we computed people's emotional variability, calculated as the within-person standard deviation of all completed emotion ratings within a measurement condition (SD; Eid & Diener, 1999). Third, we considered participants' degree of emotional instability in each measurement condition, operationalized as the root of the mean squared successive difference between two consecutive intensity scores (RMSSD; Jahng et al., 2008). Fourth, we evaluated the level of self-predictiveness or inertia of people's emotional life in each condition by estimating their person-specific PA and NA auto-regressions (AR; Kuppens, Allen, et al., 2010).

Next, to determine the split-half reliability of all emotion (dynamic) measures in each measurement condition, we evaluated Pearson correlations between variants computed from odd (i.e., 1 and 3) versus even (i.e., 2 and 4) assessment days (Wendt et al., 2020). Higher correlations indicated higher short-term consistency, and we used a Fisher's *z*-transformation to determine whether split-half correlations differed significantly from each other.

Finally, we evaluated the explanatory power (R^2) of each emotion (dynamic) measure, in each measurement condition, for both PA and NA, in the prediction of people's neuroticism level and borderline symptomatology. In a second step, we repeated these analyses, but additionally controlled for the mean level of PA or NA in that measurement condition. We obtained the relative incremental explanatory power of each emotion dynamic by fitting two series of stepwise regression models in which the order of inclusion of the two predictors (mean + emotion dynamic) was counterbalanced. We averaged the R^2s of the single and full models to decompose the total variability explained in our outcome into the two relative contributions of our individual predictors (Grömping, 2007).

Results

Differences in Observed Auto-correlation and Descriptive Correlations

Prior to our discussion of how well each measurement condition performed on the five quality criteria, we evaluated differences in observed auto-correlation. A higher auto-correlated time series may be suggestive of a response style that relies more on a relative comparison with previous emotion information, which should be evident in ABS-REL and REL measurements. The sample distributions in Figure 2 demonstrate that this was the case. For PA, ABS-REL (M = 0.64) and REL (M = 0.89) time

series exhibited stronger auto-correlation than ABS time series (M = 0.41), t's(177) ≥ 11.54, p's ≤ .001 (Panel A). An identical pattern of results was observed for NA time series (Panel B). ABS-REL (M = 0.59) and REL (M = 0.90) measurements produced higher auto-correlated ratings compared to traditional ABS assessments (M = 0.38), t's(177) ≥ 8.96, p's ≤ .001. Together, these results indicate that relative emotion ratings indeed solicit the inclusion of previous emotion levels in a momentary assessment (either visually or verbally)[5].

Next, we also examined the between-person variability in observed auto-correlations between measurement conditions. If ABS-REL and REL methods stimulated participants to use a similar reference type in their emotional judgements, this should be reflected in a more homogeneous distribution. Comparisons revealed that this was largely the case (see Figure 2). For PA, the variability in observed auto-correlations was smaller when people used ABS-REL (SD = 0.19) and REL (SD = 0.08) rating scales than when they relied on ABS measurements (SD = 0.24), F's(177,177) ≥ 1.58, p's ≤ .002. For NA, we observed significant differences between ABS (SD = 0.25) and REL (SD = 0.08), F(177,177) = 9.53, p < .001, but not between ABS and ABS-REL (SD = 0.23), F(177,177) = 1.17, p = .287. Together, these findings broadly suggest that the explicit inclusion of an emotional reference point results in the adoption of a more homogeneous response style adoption across participants (compared to unguided ABS ratings)[6].

[Figure 2 around here]

Finally, we present the Pearson correlations between (similar) emotion (dynamic) measures

[5] Despite our a priori elimination of participants that supposedly misunderstood the REL measurement method (see also Footnote 3), we still observed a considerable number of stationarity violations in this condition as a result of the drift process explained in the introduction. This non-stationarity may have inflated the observed auto-correlation in people's REL time series, which warrants caution in the interpretation of these REL results. In contrast, ABS and ABS-REL times series were equally stationary.

[6] To determine what people's natural response style was under unguided rating circumstances, we also investigated intra-individual changes in observed auto-correlation between absolute and relative measurement conditions. These analyses are presented in Supplementary Materials 5, and the estimates suggest that 10 to 37% of our participants spontaneously relied on a comparison with previous emotion information in their momentary ABS ratings. Being previously exposed to a relative measurement condition had no effect on people's adopted response style in the ABS condition (i.e., no training effects).

across and within measurement conditions in Figure 3 (see Supplementary Materials 7 and 8 for full correlation matrix and distribution differences, respectively). Contrasting the different metrics, mean levels of emotion correlated considerably across conditions for PA (r's $\geq$ 0.58) and NA (r's $\geq$ 0.47), as did variability and instability dynamics in PA (r's $\geq$ 0.45) and NA (r's $\geq$ 0.40). In contrast, emotional inertia showed the lowest cross-condition correlations for both PA (r's $\leq$ 0.23) and NA (r's $\leq$ 0.19). Contrasting the different measurement conditions, REL-based metrics consistently showed the weakest associations with those of other conditions. Finally, evaluating the mean-variability dependency in each measurement condition, weak associations were observed for PA (|r|'s $\leq$ 0.24), but not NA (|r|'s $\geq$ 0.50). This corroborates the earlier finding that problematic floor-effects are particularly apparent in negative (bounded) emotion rating scales (Kalokerinos et al., 2020; Mestdagh et al., 2018). However, contrary to our hypotheses, unbounded REL ratings did not dissolve this mean-variability confound. Unexpectedly, we observed a strong negative association between people's mean and variability in this measurement condition (r = -0.51). Investigating this further, we found that this inverse correlation was a consequence of the highlighted drift process in relative rating scales, described in our introduction (see Supplementary Materials 1 for a more in-depth discussion of this phenomenon and its implications, together with some simulated and empirical graphical examples).

[Figure 3 around here]

User Experience

Figure 4 visualizes the proportion of participants that selected a particular measurement condition as most applicable for each usability criterion. For the total UX score, proportion z-tests indicated that people significantly preferred ABS-REL over traditional ABS measurements to rate their emotions (54% vs 40%, z = 5.12, p < .001). This was particularly true for the perceived ease with which participants could rate their emotions (58% vs 37%), their self-reported reliability (60% vs 33%) and accuracy (62% vs 32%), and their willingness to re-use ABS-REL ratings in future ESM studies (64% vs 32%, z's $\geq$ 3.04, $p \leq$ .002). For perceived quickness, no significant difference were observed (48% vs 45%, z = 0.50, p = .646). However, this impression did not correspond with people's objective response times, where the average speed of survey completion was significantly faster in the ABS condition (although the effect was negligible in practical sense with only 1 second difference for the entire survey;

see Supplementary Materials 4). Next, in terms of naturalness, participants significantly preferred ABS over ABS-REL ratings (61% vs 32%, $z = 4.12$, $p < .001$). Finally, the proportion of participants that selected REL as best fitting method was the lowest for all criteria (≤ 7%).

[Figure 4 around here]

Internal Validity of ESM Time Series

Figure 5 visualizes the sample distributions of the measurement error variance in people's PA and NA time series for all three measurement conditions. For PA time series (Panel A), paired samples t-tests revealed significant differences between all measurement conditions, with error estimates being significantly lower for ABS-REL ratings ($M = 48.81$, $SD = 54.09$) and REL ratings ($M = 61.47$, $SD = 71.48$) compared to ABS ratings ($M = 83.10$, $SD = 59.42$), t's$(177) \geq 4.10$, p's $< .001$. This was partly replicated in NA time series (Panel B), where measurement error variance in ABS-REL ratings ($M = 35.26$, $SD = 47.81$), $t(177) = 9.33$, $p < .001$, but not REL ratings ($M = 66.76$, $SD = 81.32$), $t(177) = 0.07$, $p = .942$, was significantly lower than ABS ratings ($M = 67.17$, $SD = 57.47$). Finally, ABS-REL time series showed significantly less measurement error variance than REL time series, for both PA and NA, t's$(177) \geq 2.49$, $p \leq .014$. In sum, these findings consistently illustrate that measurement precision is highest when relying on ABS-REL rating scales.

[Figure 5 around here]

External Validity of ESM Time Series

First, regarding the reliability of all emotion (dynamic) measures in each measurement condition, Figure 6 visualizes the split-half correlation as a measure of short-term consistency (see Supplementary Materials 9 for corresponding numerical values and p-values of all tested differences). For emotion (dynamic) measures based on ABS ratings, our results largely corroborate the earlier finding that reliability indices decrease for measures that are more complex (Wendt et al., 2020). Specifically, we observed the highest split-half reliabilities for mean levels of PA and NA (r's ≥ 0.74, p's $_{\text{M vs SD}} \leq .012$, p's $_{\text{M vs RMSSD}} \leq .020$), followed equally by variability (r's ≥ 0.63) and instability in PA and NA (r's ≥ 0.62, p's $_{\text{SD vs RMSSD}} \geq .785$), and the lowest reliabilities for inertia in PA and NA (r's ≤ 0.23, p's $_{\text{AR vs SD}} < .001$, p's $_{\text{AR vs MSSD}} < .001$). However, critical to our research question, this cascade was countered in the REL condition for emotional instability. Compared to ABS, the split-half reliability

of people's PA and NA instability was significantly higher in a REL context (r's ≥ 0.79, p's $_{\text{REL vs ABS}} \leq .001$). This suggests that a direct assessment of this emotion dynamic can improve its short-term consistency. In a similar vein, for emotional inertia, the reliability was higher when relying on ABS-REL ratings in NA ($r = 0.32$, $p_{\text{ABS vs ABS-REL}} = .010$), but not PA ($r = 0.24$, $p_{\text{ABS vs ABS-REL}} = .908$). Nevertheless, in absolute terms, the reliabilities for emotional inertia were still unacceptably low, which questions the stable between-person quality of this emotion dynamic.

For mean levels of emotion, REL ratings produced significantly lower reliabilities than ABS (p's $_{\text{REL vs ABS}} \leq .015$). In contrast, ABS versus ABS-REL did not produce significant differences in the reliability of mean levels of emotion (p's $_{\text{ABS vs ABS-REL}} \geq .247$). Finally, regarding the split-half reliability of emotional variability, we observed no significant differences between measurement conditions (p's $_{\text{ABS vs ABS-REL vs REL}} \geq .199$).

[Figure 6 around here]

Second, in terms of univariate explanatory power, Figure 7 visualizes the R^2s of all emotion (dynamic) measures in each measurement condition (see Supplementary Materials 10 for corresponding numerical values). Contrasting the different measures (irrespective of measurement condition), mean levels of PA and NA generally showed the strongest association with neuroticism (average $R^2 = 0.26$), while borderline symptomatology was predicted more or less equally well by people's means (average $R^2 = 0.18$) and variability in PA and NA (average $R^2 = 0.15$). This replicates earlier work that the explanatory power of more complex measures is rather low compared to these more basic summaries (average $R^2_{\text{RMSSD}} = 0.12$, average $R^2_{\text{AR}} = 0.01$; Dejonckheere et al., 2019; Wendt et al., 2020). Focal to our research question (and disconfirming our hypotheses), the use of relative ratings did not result in any stronger explanatory effects of emotion dynamic measures in the prediction of these person-level outcomes. Specifically, ABS-REL ratings did not substantially improve emotional inertia's prediction of neuroticism profiles in absolute terms (R^2s ≤ 0.02). Similarly, REL ratings did not produce higher effect sizes for variability and instability in the prediction of neuroticism (R^2s ≤ 0.04) or borderline symptoms (R^2s ≤ 0.07). In fact, compared to ABS and ABS-REL measurements, REL-based summaries generally performed poorly in the prediction of external criteria. Unexpectedly, the use of ABS-REL ratings more or less doubled the explanatory power of emotional variability in the prediction of people's

neuroticism scores, for both PA (R^2_{ABS} = 0.07 vs $R^2_{ABS\text{-}REL}$= 0.14) and NA (R^2_{ABS} = 0.15 vs $R^2_{ABS\text{-}REL}$= 0.28). Taken together, besides emotional variability, relying on alternative emotion measurements does not meaningfully improve the individual predictive ability of emotion dynamic measures.

[Figure 7 around here]

Third, in terms of added explanatory power above people's mean level of positive or negative emotion, Figure 8 presents the relative contribution in R^2 of a particular mean and emotion dynamic combination in each measurement condition, in the prediction of people's neuroticism or borderline personality scores (see Supplementary Materials 10 for corresponding numerical values). In line with previous (ABS-based) studies (Dejonckheere et al., 2019; Wendt et al., 2020), the incremental explanatory power of emotion dynamic measures was generally low compared to static mean levels of PA or NA, with a few notable exceptions. Regarding the prediction of neuroticism, the relative contribution of emotional variability in an ABS-REL context was remarkably larger than the relative contribution of an ABS-based variant, for both PA (relative R^2_{ABS} = 0.04 vs relative $R^2_{ABS\text{-}REL}$ = 0.11) and NA (relative R^2_{ABS} = 0.08 vs relative $R^2_{ABS\text{-}REL}$ = 0.17). Although we did not hypothesize this a priori, this finding suggests that neuroticism could be characterized by stronger emotional variability along differences in basic mean levels of positive and/or negative emotion, but that this depends on the type of measurement that is used in ESM (Kalokerinos et al., 2020). Similarly, in the prediction of borderline symptoms, the relative R^2s of emotional variability and instability more or less equal those of people's mean emotion levels in an ABS-REL context for PA (relative R^2s $_{M} \leq .18$ vs relative R^2s $_{SD\text{ and }RMSSD} \leq .15$), and outperform the relative R^2s of mean levels in NA (relative R^2s $_{M} \leq .11$ vs relative R^2s $_{SD\text{ and }RMSSD} \leq .17$). Finally, in terms of total R^2s, we did not observe notable differences between models based on ABS (average R^2 = 0.31) and ABS-REL (average R^2 = 0.30) predictors. In contrast, the performance of REL-based models was consistently poorer (average R^2 = 0.20).

[Figure 8 around here]

General Discussion

The goal of this study was to experimentally evaluate to what extent relative emotion assessments could overcome some of the theoretical and methodological challenges of traditional absolute emotion ratings in ESM. Either encouraging an implicit comparison with one's previous

emotion rating through a relative 'Last' anchor on an absolute measurement scale (ABS-REL), or explicitly soliciting an emotion comparison via relative item phrasings and ditto rating scales (REL), we aimed to increase participants' rating experience, as well as the internal and external validity of the resulting emotion time series. In the next paragraphs, we evaluate the performance of each assessment method, and formulate specific recommendations for future ESM research on emotions.

Absolute Emotion Ratings With and Without Relative 'Last' Anchor (ABS vs ABS-REL)

Theoretically, we found that the visualization of participants' previous emotional response on the absolute emotion intensity scale improved their overall user experience. Specifically, the general preference to re-use this type of scale in future ESM studies and its perceived easiness highlight the superior face validity of this assessment procedure (Nevo, 1985). It is possible that this is because participants felt that this type of measurement better aligned with their own conceptualization of emotions. If so, our UX findings fit within the *adaption level* literature on emotional judgements (Helson, 1964; Russell & Lanius, 1984), which demonstrates that people's current emotional experience is (partly) shaped by a relative comparison with previous emotional information (e.g., Kahneman, 1992). The addition of a 'Last' anchor was also deemed more effective with respect to the perceived accuracy and reliability of people's momentary emotion ratings. This could indicate that the inclusion of a clear reference point enables a more structured interpretation of the measurement scale (Yannakakis et al., 2017), or that it allows for a more precise evaluation of emotional fluctuations. Regardless of the exact explanatory mechanism, reduced between-person differences in the increased auto-correlation of participants' ABS-REL time series suggest that they all more include previous emotion information in their momentary ratings, introducing more standardization in rating behavior across individuals (as opposed to free response styles in ABS ratings where participants can are not guided in the reference value they adopt, see Figure 2). Finally, from the researcher's perspective, another UX advantage is that the anchor may facilitate a successful relative emotional evaluation even if participants occasionally miss or skip an assessment (if the anchor does not refer to an emotional assessment that lies too far in the past, see Footnote 4). Providing information about the exact timing of the previous assessment could further aid participants in their recall and associated relative comparison.

Methodologically, these endorsed superior qualities of ABS-REL measurements resonated in a

higher internal validity of participants' emotion time series (Dejonckheere & Mestdagh, 2021). That is, for both positive and negative momentary emotional experiences, the explicit presentation of previous emotion levels resulted in a lower estimated measurement error variance in people's time series (Schuurman & Hamaker, 2019). This may be because the annotation of a 'Last' anchor equipped participants with an accurate pointer to indicate similar intensity scores when their true emotion levels had not meaningfully changed across assessments. More measurement precision resulted in a better representation of people's true (within-person) emotional ups and downs.

Consequently, in terms of external validity, these higher-quality ABS-REL time series produced person-level emotion dynamic summaries that were more reliable in certain instances. While the short-term consistency of simpler metrics, like people's emotional mean or variability, was equally satisfactory when relying on traditional ABS versus alternative ABS-REL ratings, the split-half reliability for their negative, but not positive, emotional inertia was more than six time higher in the latter measurement context. Still, in absolute terms, this increase was not sufficient to justify the recognition of emotional inertia as a stable between-person construct (Wendt et al., 2020). Consequently, when considering the individual and added explanatory power of this dynamic in the prediction of people's neuroticism (but also borderline symptoms), the implementation of a 'Last' anchor did not significantly improve its effect sizes. Together, these findings are more suggestive of the minimal role of emotional inertia in people's personality or well-being, rather than that certain measurement practices in traditional ESM would conceal any meaningful relations (Dejonckheere et al., 2019).

This stands in stark contrast to the significant role of emotional variability (and instability) in neuroticism (but also borderline symptoms). In terms of univariate explanatory power, a simple reference to one's previous emotion score on the absolute measurement scale unexpectedly doubled the effect sizes of positive and negative emotional variability in the prediction of people's neuroticism profiles (an increase of 8 and 14%, respectively). In terms of added explanatory power above people's static mean levels of PA or NA, this translated into global effect sizes that were better represented by the relative contribution of emotional variability. These findings illustrate that the emotional life of neurotic individuals could be characterized by more intense fluctuations, next to how positive and

negative they feel on average, but that this depends on how their emotions were assessed during ESM (Kalokerinos et al., 2020). In hindsight, (a combination of) multiple accounts could explain the superiority of an ABS-REL framework to assess between-person differences in emotional variability in ESM: Compared to ABS, (1) the reduced between-person heterogeneity in response styles may result in better comparable time series summaries, (2) the stronger focus on relative changes in emotion over time gives a more equal weight to the first (M) and second (SD) statistical moment of a group of emotion scores, and (3) the relative anchor results in more measurement precision at the within-person level, which also positively impacts the external validity of a time series.

Absolute versus Relative Item Phrasing (ABS vs REL)

In contrast, phrasing momentary emotion assessments in a relative way was generally unsuccessful (even after the removal of some inaccurate subjects). For instance, regarding their perceived user experience, participants consistently nominated this type of measurement as the least preferred method for all UX criteria. The reason for this depreciation may be that a self-reliant recollection of previous emotion levels was too difficult (Retkoceri, 2022), discouragingly inaccurate due to emotional recall biases (e.g., Robinson & Clore, 2002), or cognitively too burdensome for people to repeatedly access this information from their explicit memory (Carlsson, 1983; Sanna & Chang, 2006). Indeed, the two-step approach with which participants first had to remember how they felt earlier, in order to then evaluate how they deviated from this reference point may have hindered a natural and effortless assessment of people's momentary emotions (e.g., Nevo, 1985). Furthermore, when ESM assessments are frequently missed or skipped, the time interval between REL ratings likely becomes longer. This could undermine the perceived functionality of this measurement method even further.

Despite participants' unfavorable judgment, these relatively phrased items did produce ESM time series that generally exhibited fewer measurement error. However, these results should be interpreted with caution, because stationarity violations in people's (absolute-transformed) REL time series, likely inflated the (latent) auto-regressive estimates (Bringmann et al., 2013). In contrast, differences in the measurement error variance between absolute versus relative emotion time series were less outspoken (only a significant reduction for PA). This may be due to the fact that the scale midpoint likely introduced more measurement precision to rate instances of emotional stability, but only to an

extent that participants were able to accurately recall their previous emotional state (Retkoceri, 2022). More worrisome, for instances of meaningful emotional change, the inevitable measurement error associated with each rating accumulated over time, because people's previously distorted emotion rating served as a (mental) departure level for people's current emotion rating, which again showed some degree of measurement error that carried over to the next assessment. The problematic resonation of this error term is exactly the reason why we observed an unusual drift process in some people's observed (absolute-transformed) relative emotion ratings, and why the NA mean-variability association unexpectedly reversed sign in a relative framework (see Supplementary Materials 1 for some visual examples)[7].

With respect to the external validity of relative emotion time series (Dejonckheere & Mestdagh, 2021), we observed how responses to relatively phrased emotion items produced less reliable estimations of people's average PA and NA. In contrast, the split-half correlations for people's positive and negative emotional instability were significantly higher in a relative context. The reason for this pattern is that a comparative evaluation of people's current emotional intensity in relation to their previous emotion level directly corresponds to the theoretical conceptualization of emotional instability (Jahng et al., 2008). As opposed to traditional absolute emotion ratings in ESM, this straightforward assessment does not require complex computations of the data (i.e., one simply needs to take the mean of people's REL ratings to obtain their emotional instability), which positively impacts the short-term stability of this index (Wendt et al., 2020).

Even though relatively phrased items improved the reliability of emotional instability, this did not translate into stronger effect sizes for this dynamic in the prediction of borderline personality symptoms (or neuroticism). In fact, the individual predictive effects of REL-based (dynamic) metrics were consistently lower than those of traditional ABS-based measures. Similarly, regarding the incremental explanatory power above mean levels of PA or NA, we found that REL-based models could not outpredict traditional ABS-based models. A possible explanation for this poor predictive

[7] This also explains why REL-based metrics generally showed weaker association with similar emotion (dynamic) measures computed from ABS or ABS-REL time series.

performance is that, contrary to our expectations, the use of a relative measurement scale did not resolve the critical mean-variability dependency seen in (negative) ABS-based emotion time series (Kalokerinos et al., 2020; Mestdagh et al., 2018). While a floor-effect in NA ratings created a strong positive association between people's mean and variability in an ABS context, the sign of this relation flipped in a REL context (leaving the size of this relation unchanged). Because most emotion dynamic measures therefore showed considerable overlap with people's mean level of emotion, their relative contribution in the prediction of external criteria was still limited.

Recommendations for Future ESM Research on Emotions

Based on these criteria, we see value in the use of ABS-REL measurements in emotional ESM, because they combine the advantages of both a relative comparison and an absolute rating scale. For instance, the 'Last' marker provides a standardized and tangible reference point to structure the ambiguous rating space within and across individuals (as opposed to ABS), but does not rely on an accurate retrieval of previous emotion levels that may be too demanding (as opposed to REL). Furthermore, the relative evaluation that is induced by a 'Last' anchor generates less between-person heterogeneity in response style (as opposed to ABS), while the absolute properties of the scale prohibit the problematic accumulation of measurement error over time (as opposed to REL). Moreover, although ABS-REL measurements do not dissolve the inconvenient mean-variability confound in NA time series (similar to ABS and REL), they do reveal the individual and added predictive effect of emotional variability (and instability) in people's neuroticism scores (as opposed to ABS and REL). Finally, participants generally nominate this type of measurement as the best method in terms of user experience.

Constraints on Generality and Limitations

The present findings should be interpreted in the light of some limitations. First, regarding the internal validity of people's emotion time series, it is worth emphasizing that the true model underlying people's emotional life can never be defined. Although affective researchers frequently use the latent AR(1) model to describe the dynamic regularities of

people's true emotions (Kuppens & Verduyn, 2017), different model operationalizations will produce different measurement error estimates (Rhemtulla et al., 2020). In this regard, a common criticism of the latent AR(1) model is that it is too basic to accurately represent the complex nature of our emotional system (Loossens et al., 2020). This generally leads to an overestimation of measurement error variance (Dejonckheere et al., 2022), but it remains unclear whether different types of measurement are affected to a similar extent. Relatedly, when model assumptions of the latent AR(1) model are severely violated (e.g., non-stationarity in REL time series; Bringmann et al., 2013), the biased parameters will not produce a reliable estimate of the measurement error variance in people's ESM time series. As mentioned in Footnote 5, the internal validity of the REL time series is likely inflated.

Second, we only established the superiority of ABS-REL ratings in a specific context. That is, we limited our analyses to the momentary assessment of emotions, using visual slider scales in a relatively healthy sample. Whether the addition of a 'Last' anchor in ESM is also beneficial for a more accurate assessment of other subjective experiences (e.g., momentary psychological complaints), when relying on shorter or labeled measurement scales (e.g., 5-point Likert scales with qualitative intensity indicators), in different study populations (e.g., psychiatric patients who tend to report higher levels of negative emotion), remains to be investigated. Similarly, the possibility exists that in the light of other, uninvestigated quality criteria in ESM, this measurement method may not outperform traditional absolute intensity scales. Overall, future research could benefit from a more fine-grained analysis that explores what works best for whom, and under what circumstances (e.g., different types of research questions may solicit different assessment methods).

Finally, to further unravel the favorable working mechanisms of ABS-REL ratings, it could be worthwhile to explicitly test the assumption that it is the solicitation to incorporate previous emotion levels in participants' current assessment that explains its better performance.

Comparing our results with a measurement condition in which we display incorrect 'Last' anchors on the measurement scale could provide a definite answer to this question. Related to this issue, it remains an open question what the status of the first assessment in an ABS-REL framework is, as this rating cannot be based on a relative comparison with (meaningful) previous emotion information. Depending on whether researchers aim to pursue full assessment standardization within and between individuals, they could decide to include this first rating in further analyses or not.

Conclusion

Based on these five quality criteria that are important for ESM, we believe that the implementation of a visual 'Last' anchor could improve emotion measurement in daily life. Theoretically, people prefer this method to rate their emotions, resulting in a higher face validity of the assessment procedure. Methodologically, this measurement method produces time series with a higher internal and, in some cases, higher external validity.

References

Aitken, R. C. (1969). Measurement of feelings using visual analogue scales. *Proc R Soc Med*, *62*, 989–993.

Baird, B. M., Le, K., & Lucas, R. E. (2006). On the nature of intraindividual personality variability: Reliability, validity, and associations with well-being. *J Pers Soc Psychol*, *90*, 512–527.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE*, *8*, e60188.

Carlsson, A. M. (1983). Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale. *Pain*, *16*, 87–101.

Carpenter, R. W., & Trull, T. J. (2013). Components of emotion dysregulation in borderline personality disorder: A review. *Curr Psychiatry Rep*, *15*, 335.

Chow, S.-M., Ram, N., Boker, S. M., Fujita, F., & Clore, G. (2005). Emotion as a thermostat: Representing emotion regulation using a damped oscillator model. *Emotion*, *5*, 208–225.

De Vuyst, H.-J., Dejonckheere, E., Gucht, K. V. der, & Kuppens, P. (2019). Does repeatedly reporting positive or negative emotions in daily life have an impact on the level of emotional experiences and depressive symptoms over time? *PLoS ONE*, *14*, e0219121.

Dejonckheere, E., Bastian, B., Fried, E. I., Murphy, S. C., & Kuppens, P. (2017). Perceiving social pressure not to feel negative predicts depressive symptoms in daily life. *Depress Anxiety*, *34*, 836–844.

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

Dejonckheere, E., & Erbas, Y. (2021). Designing an experience sampling study. In *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (pp. 33–70). Center for research on experience sampling and ambulatory methods Leuven.

Dejonckheere, E., & Mestdagh, M. (2021). On the signal-to-noise ratio in real-life emotional time

series. In C. E. Waugh & P. Kuppens (Eds.), *Affect Dynamics* (pp. 131–152). Springer International Publishing.

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. *Nat Hum Behav*, *3*, 478–491.

Dejonckheere, E., Mestdagh, M., Kuppens, P., & Tuerlinckx, F. (2020). Reply to: Context matters for affective chronometry. *Nat Hum Behav*, *4*, 690–693.

Denissen, J., Geenen, R., Aken, M., Gosling, S., & Potter, J. (2008). Development and validation of a dutch translation of the Big Five Inventory (BFI). *J Pers Assess*, *90*, 152–157.

Distel, M. A., de Moor, M. H. M., & Boomsma, D. I. (2009). Nederlandse vertaling van de Personality Assessment Inventory - Borderline kenmerken schaal (PAI-Bor): Normgegevens, factorstructuur en betrouwbaarheid. *Psychologie en Gezondheid*, *37*, 38–46.

Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *J Pers Soc Psychol*, *76*, 662–676.

Eisele, G., Kasanova, Z., & Houben, M. (2021). Questionnaire design and evaluation. In *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (pp. 71–100). Center for research on experience sampling and ambulatory methods Leuven.

Eysenck, M. (1985). *Personality and individual differences: A natural science approach* (1st edition). Plenum Press.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *Am Stat*, *61*, 139–147.

Halladay, L. R., Zelikowsky, M., Blair, H. T., & Fanselow, M. S. (2012). Reinstatement of extinguished fear by an unextinguished conditional stimulus. *Front Behav Neurosci*, *6*, 18.

Haslam, N. (2017). Bell-shaped distribution of personality traits. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 1–2). Springer International Publishing.

Heininga, V. E., Dejonckheere, E., Houben, M., Obbels, J., Sienaert, P., Leroy, B., van Roy, J., &

Kuppens, P. (2019). The dynamical signature of anhedonia in major depressive disorder: Positive emotion dynamics, reactivity, and recovery. *BMC Psychiatry*, *19*, 59.

Helson, H. (1947). Adaptation-level as frame of reference for prediction of psychophysical data. *American J Psychol*, *60*, 1–29.

Helson, H. (1964). *Adaptation-level theory: An experimental and systematic approach to behavior*. NY.

Hisler, G. C., Krizan, Z., DeHart, T., & Wright, A. G. C. (2020). Neuroticism as the intensity, reactivity, and variability in day-to-day affect. *J Res Pers*, *87*, 103964.

Houben, M., Mestdagh, M., Dejonckheere, E., Obbels, J., Sienaert, P., van Roy, J., & Kuppens, P. (2021). The statistical specificity of emotion dynamics in borderline personality disorder. *J Pers Disord*, *35*, 819–840.

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychol Bull*, *141*, 901–930.

Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychol Methods*, *13*, 354–375.

Jaso, B. A., Kraus, N. I., & Heller, A. S. (2021). Identification of careless responding in ecological momentary assessment research: From posthoc analyses to real-time data monitoring. *Psychol Methods. 27*, 958–981.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research 2nd ed* (pp. 102–138). Guilford Press.

Kahneman, D. (1992). Reference points, anchors, norms, and mixed feelings. *Organ Behav Hum Decis Process*, *51*, 296–312.

Kalokerinos, E. K., Murphy, S. C., Koval, P., Bailen, N. H., Crombez, G., Hollenstein, T., Gleeson, J., Thompson, R. J., Ryckeghem, D. M. L. V., Kuppens, P., & Bastian, B. (2020). Neuroticism may not reflect emotional variability. *Proc Natl Acad Sci U S A*, *117*, 9270–9276.

Kalokerinos, 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. *Psychol Sci*, *30*, 863–879.

Koval, P., Pe, M. L., Meers, K., & Kuppens, P. (2013). Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion*, *13*, 1132–1141.

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychol Sci*, *21*, 7.

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *J Pers Soc Psychol*, *99*, 1042–1060.

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Curr Opin Psychol*, *17*, 22–26.

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. *New Dir Teach Learn*, *15*, 41–56.

Larson, R., Csikszentmihalyi, M., & Graef, R. (1980). Mood variability and the psychosocial adjustment of adolescents. *J Youth Adolesc*, *9*, 469–490.

LeDoux, J. E. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. Simon & Schuster.

Levine, L. J., Prohaska, V., Burgess, S. L., Rice, J. A., & Laulhere, T. M. (2001). Remembering past emotions: The role of current appraisals. *Cogn Emot*, 15, 393–417.

Lockhead, G. R. (2004). Absolute judgments are relative: A reinterpretation of some psychophysical ideas. *Rev Gen Psychol*, *8*, 265–272.

Loossens, T., Mestdagh, M., Dejonckheere, E., Kuppens, P., Tuerlinckx, F., & Verdonck, S. (2020). The affective ising model: A computational account of human affect dynamics. *PLoS Comp Biol*, *16*, e1007860.

Mejía, S., Hooker, K., Ram, N., Pham, T., & Metoyer, R. (2014). Capturing intraindividual variation and covariation constructs: Using multiple time-scales to assess construct reliability and construct stability. *Res Hum Dev*, *11*, 91–107.

Mestdagh, M., & Dejonckheere, E. (2021). Ambulatory assessment in psychopathology research: Current achievements and future ambitions. *Curr Opin Psychol*, *41*, 1–8.

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. *Psychol Methods*, *23*, 690–707.

Mestdagh, M. , Verdonck, S., Piot, M., Niemeijer, K., Tuerlinckx, F., Kuppens, P., & Dejonckheere, E. (2023). m-Path: An easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioural research and clinical practice. *Frontiers in Digital Health*, *5*, 1182175.

Miller, D. J., Vachon, D. D., & Lynam, D. R. (2009). Neuroticism, negative affect, and negative affect instability: Establishing convergent and discriminant validity using ecological momentary assessment. *Pers Individ Differ*, *47*, 873–877.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol Rev*, *63*, 81–97.

Morey, L. (1991). *The Personality assessment inventory: Professional manual*. Psychological Assessment Resources.

Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the Middle: The use and interpretation of mid-points in items on questionnaires. *J Gen Psychol*, *142*, 71–89.

Nevo, B. (1985). Face validity revisited. *J Educ Meas*, *22*, 287–293.

Panaite, V., Koval, P., Dejonckheere, E., & Kuppens, P. (2019). Emotion regulation and mood brightening in daily life vary with depressive symptom levels. *Cogn Emot*, *33*, 1291–1301.

Parsons, V. L. (2017). Stratified sampling. In *Wiley Statistics Reference Online* (pp. 1–11). John Wiley & Sons, Ltd.

Pe, M., 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. *Clin Psychol Sci*, *3*, 292–300.

Ram, N., Brinberg, M., Pincus, A. L., & Conroy, D. E. (2017). The questionable ecological validity of ecological momentary assessment: Considerations for design and analysis. *Res Hum Dev*, *14*, 253–270.

Retkoceri, U. (2022). Remembering emotions. *Biol Philo*, *37*, 5.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychol Methods*, *25*, 30–45.

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychol Bull, 128*, 934–960.

Russell, J. A. (1980). A circumplex model of affect. *J Pers Soc Psychol, 39*, 1161–1178.

Russell, J. A., & Lanius, U. F. (1984). Adaptation level and the affective appraisal of environments. *Journal Environ Psychol, 4*, 119–135.

Sanna, L. J., & Chang, E. C. (2006). *Judgments over time: The interplay of thoughts, feelings, and behaviors.* Oxford University Press.

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychol Methods, 24*, 70–91.

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n = 1 psychological autoregressive modeling. *Front Psychol, 6*, 1038.

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proc Nat Acad Sci U S A, 115*, 15–23.

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychol Assess, 31*, 557–566.

Sims, T., Tsai, J. L., Jiang, D., Wang, Y., Fung, H. H., & Zhang, X. (2015). Wanting to maximize the positive and minimize the negative: Implications for mixed affective experience in american and chinese contexts. *J Pers Soc Psychol, 109*, 292–315.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychol Rev, 112*, 881–911.

Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Pers Soc Psychol Bull, 24*, 127–136.

Svensson, E. (2000). Concordance between ratings using different scales for the same variable. *Stat Med, 19,* 3483–3496.

Tackett, J. L., & Lahey, B. B. (2017). Neuroticism. In *The Oxford handbook of the Five Factor Model* (pp. 39–56). Oxford University Press.

Thompson, R. J., Kircanski, K., & Gotlib, I. H. (2016). The grass is not as green as you think: Affect evaluation in people with internalizing disorders. *J Affect Disord*, *203*, 233–240.

Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Gotlib, I. H. (2012). The everyday emotional experience of adults with major depressive disorder: Examining emotional instability, inertia, and reactivity. *J Abnorm Psychol*, *121*, 819–829.

Trampe, D., Quoidbach, J., & Taquet, M. (2015). Emotions in everyday life. *PloS ONE*, *10*, e0145450.

Tsai, J. L. (2017). Ideal affect in daily life: Implications for affective experience, health, and social behavior. *Curr Opin Psychol*, *17*, 118–128.

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. *J Med Internet Res*, *21*, e14475.

Van der Gucht, K., Dejonckheere, E., Erbas, Y., Takano, K., Vandemoortele, M., Maex, E., Raes, F., & Kuppens, P. (2019). An experience sampling study examining the potential impact of a mindfulness-based intervention on emotion differentiation. *Emotion*, *19*, 123–131.

Vanbelle, S., & Lesaffre, E. (2018). Modeling agreement on bounded scales. *Stat Methods Med Res*, *27*, 3460–3477.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behav Res*, *43*, 476–496.

Waugh, C. E., & Kuppens, P. (2021). *Affect Dynamics*. Springer.

Wendt, L. P., Wright, A. G. C., Pilkonis, P. A., Woods, W. C., Denissen, J. J. A., Kühnel, A., & Zimmermann, J. (2020). indicators of affect dynamics: Structure, reliability, and personality correlates. *Eur J Pers*, *34*, 1060–1072.

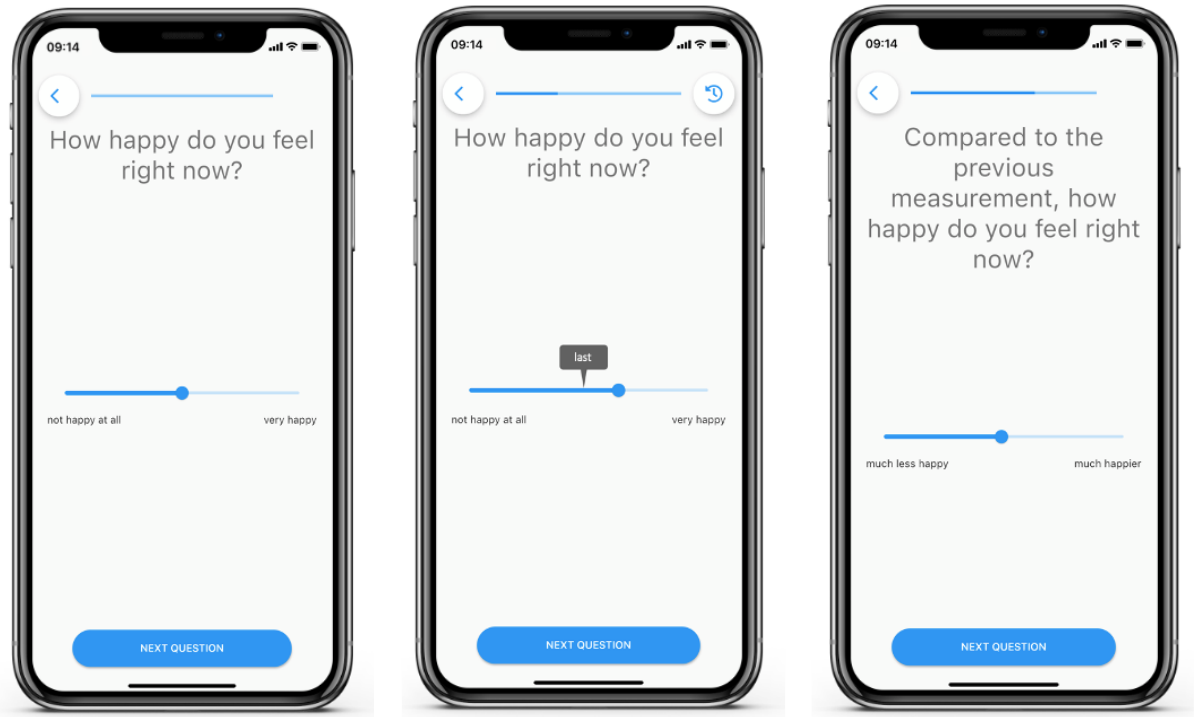
Yannakakis, G. N., Cowie, R., & Busso, C. (2017). The ordinal nature of emotions. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction*, 248–255.

Yannakakis, G. N., Cowie, R., & Busso, C. (2021). The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing*, *12*, 16–35.

Yzerbyt, V., Dumont, M., Mathieu, B., Gordijn, E., & Wigboldus, D. (2006). Social comparison and group-based emotions. In *Social comparison and social psychology: Understanding cognition,*

intergroup relations, and culture (pp. 174–205). Cambridge University Press.

Figures



A. B. C.

Figure 1. m-Path rating interface to assess participants' momentary emotion (e.g., happiness) levels in each measurement condition. Panel A depicts the absolute measurement condition (ABS; instruction: *Today, you will evaluate your current emotions in an absolute way. You will rate the intensity of your emotions on a scale from* 'not at all [emotion]' *to* 'very [emotion]'.). Panel B depicts the absolute measurement condition with relative 'Last' anchor (ABS-REL; instruction: *Today, you will evaluate your current emotions in an absolute way. In addition, the anchor point on the scale displays your emotional rating from the previous assessment occasion. You will rate the intensity of your emotions on a scale from* 'not at all [emotion]' *to* 'very [emotion]', *and you can use the anchor point 'Last' as a basis for comparison.*). Panel C depicts the relative measurement condition (REL; instruction: *Today, you will evaluate your current emotions in a relative way. From the next assessment occasion, you will compare your current emotion level with the level of that emotion from the previous assessment occasion. You will use a scale from* 'much less [emotion]' *to* 'much more [emotion]'.).

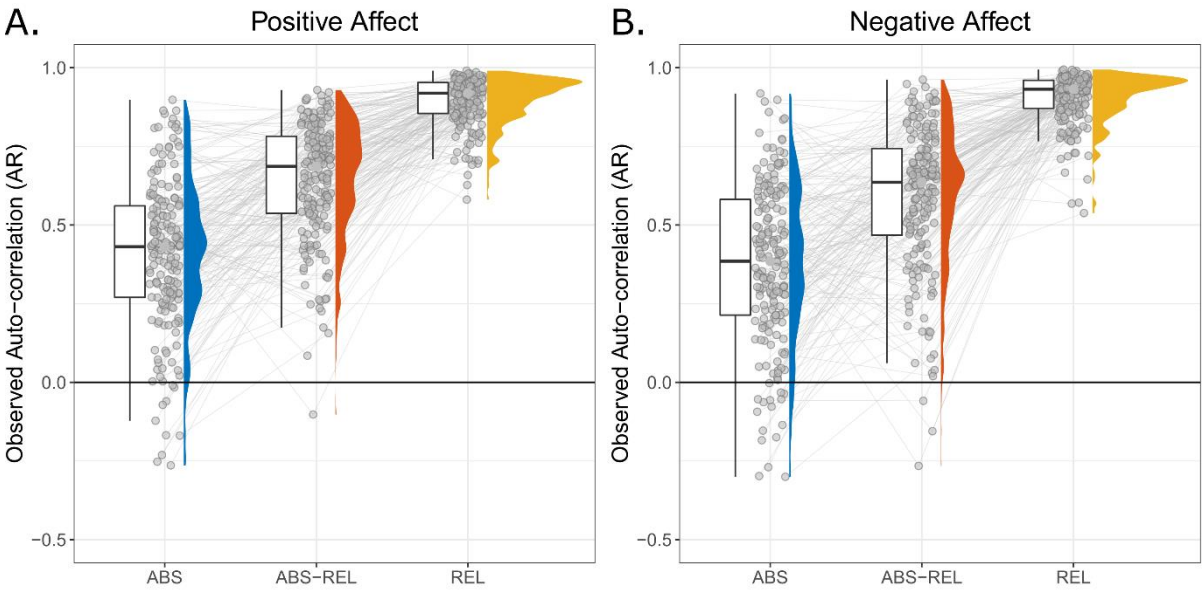


Figure 2. Investigating differences in observed auto-correlation between measurement conditions (ABS, ABS-REL and REL) for positive (left) and negative (right) affect. The graph depicts the between-person sample distributions and associated boxplots. Identical observations are connected across conditions to visualize within-person differences (n = 178).

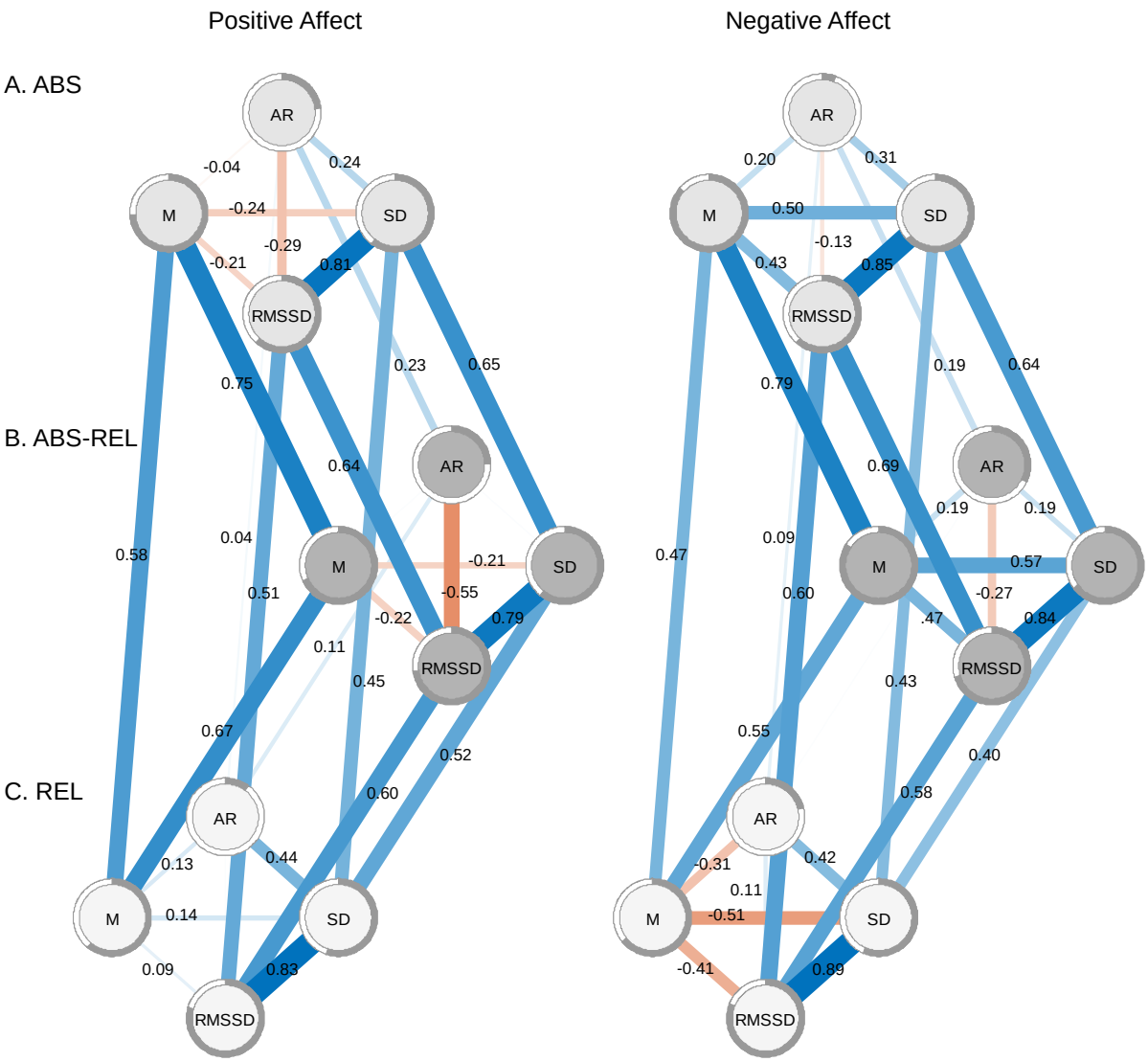


Figure 3. Investigating the empirical interrelations (edges) and split-half reliabilities (pie graphs) for all emotion (dynamic) measures (i.e., mean, variability, instability and inertia) in each measurement condition (ABS, ABS-REL and REL), for positive (left) and negative (right) affect (n =178). Blue and red edges represent positive and negative Pearson correlations, respectively. Transparency and thickness of the edges correspond with the degree of association. For clarity, interrelations between valences are not visualized (see Supplementary Materials 7 for full correlation matrix). A fuller pie graph around each node indicates a higher split-half Pearson correlation for that particular emotion (dynamic) measure (see also Figure 6).

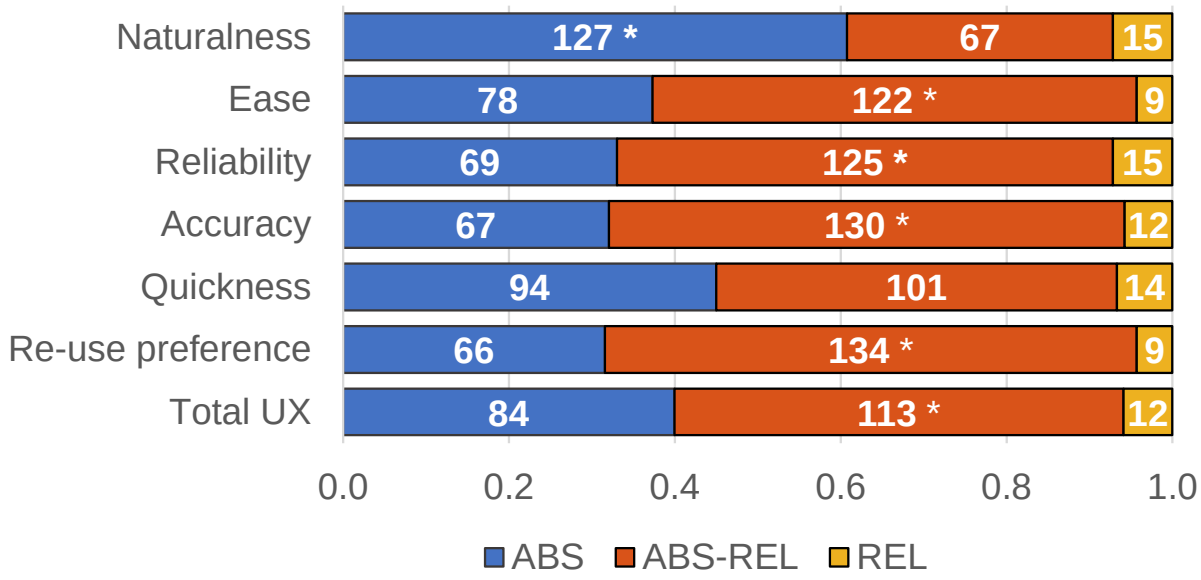


Figure 4. Evaluating people's user experience following six criteria (n = 209). The stacked bars visualize the proportion of participants that ranked a particular measurement method (ABS, ABS-REL, REL) as best fitting for that criterium. The asterisk (*) indicates a significant difference between the first versus second most selected measurement condition. The full items can be found in Supplementary Materials 3.

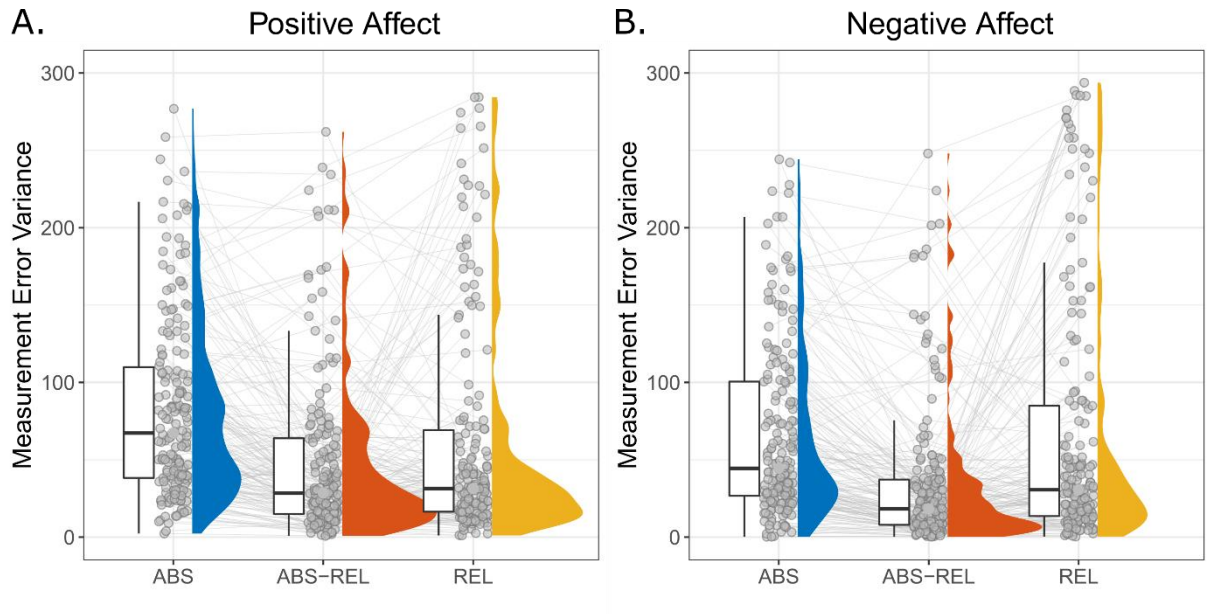


Figure 5. Evaluating the internal validity of people's emotion time series following each measurement condition (ABS, ABS-REL and REL). The graph depicts the sample distributions and associated boxplots for people's measurement error variance, for positive (left) and negative (right) affect. Identical observations are connected across conditions to visualize within-person differences (n = 178).

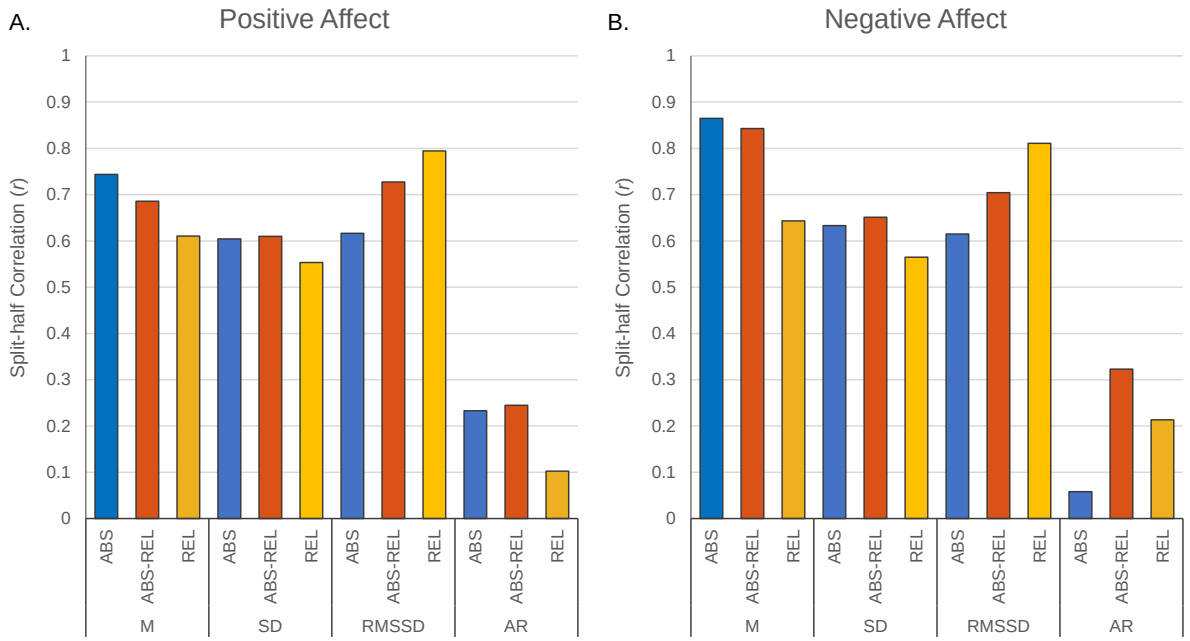


Figure 6. Investigating the split-half reliabilities for all emotion (dynamic) measures (i.e., mean, variability, instability and inertia) in each measurement condition (ABS, ABS-REL and REL), for positive (left) and negative (right) affect (n =178; see Supplementary Materials 8 for corresponding numerical values and p-values of comparison tests).

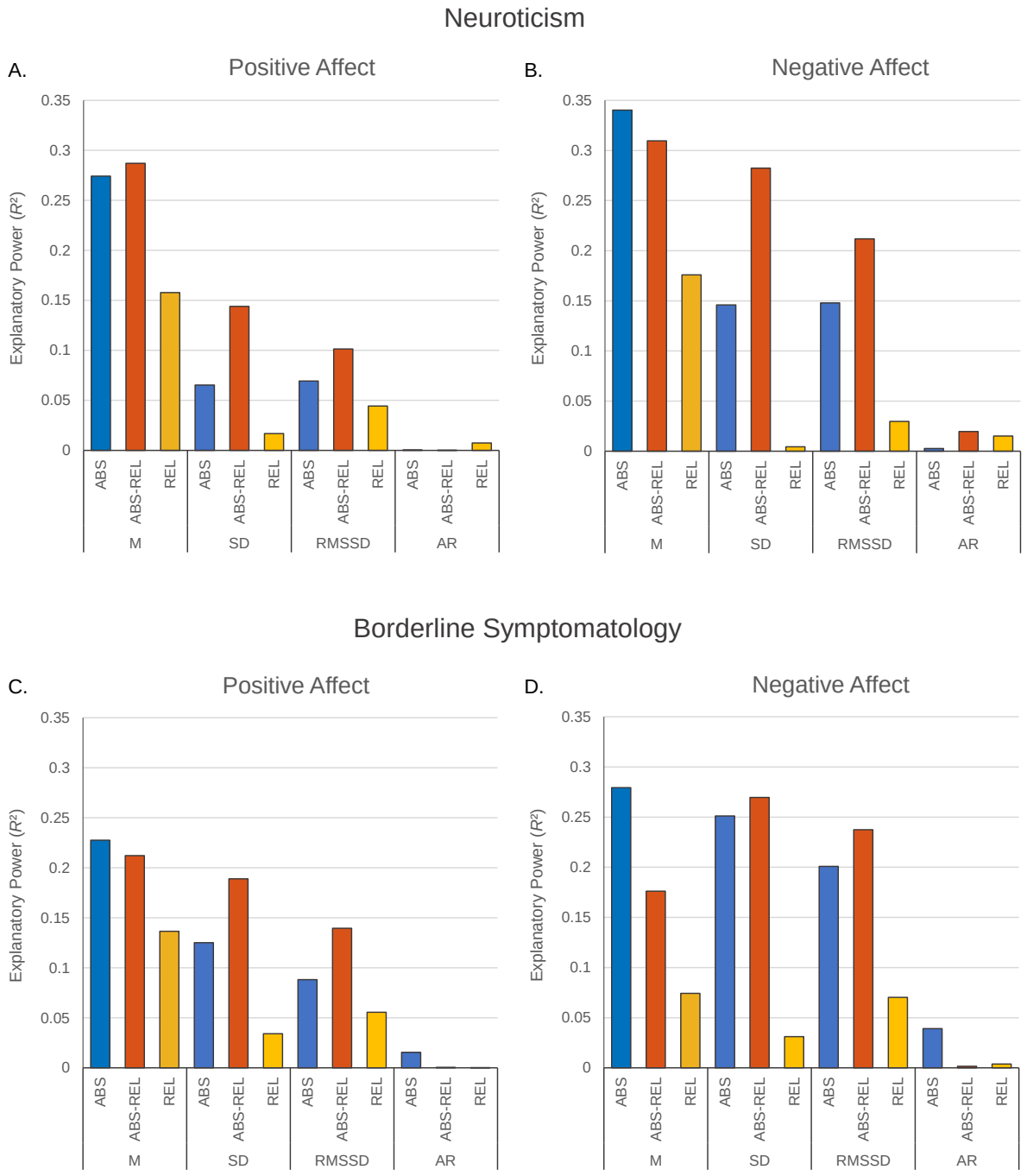


Figure 7. Evaluating the individual external validity of people's emotion time series following each measurement condition. The two-by-two graph depicts the explanatory power (R^2) for each emotion (dynamic) measure (i.e., mean, variability, instability and inertia) computed from a different time series type (ABS, ABS-REL, REL) in the prediction of people's neuroticism scores (top; n = 176) and borderline symptomatology (bottom; 171), for positive (left) and negative (right) affect (see Supplementary Materials 10 for corresponding numerical values).

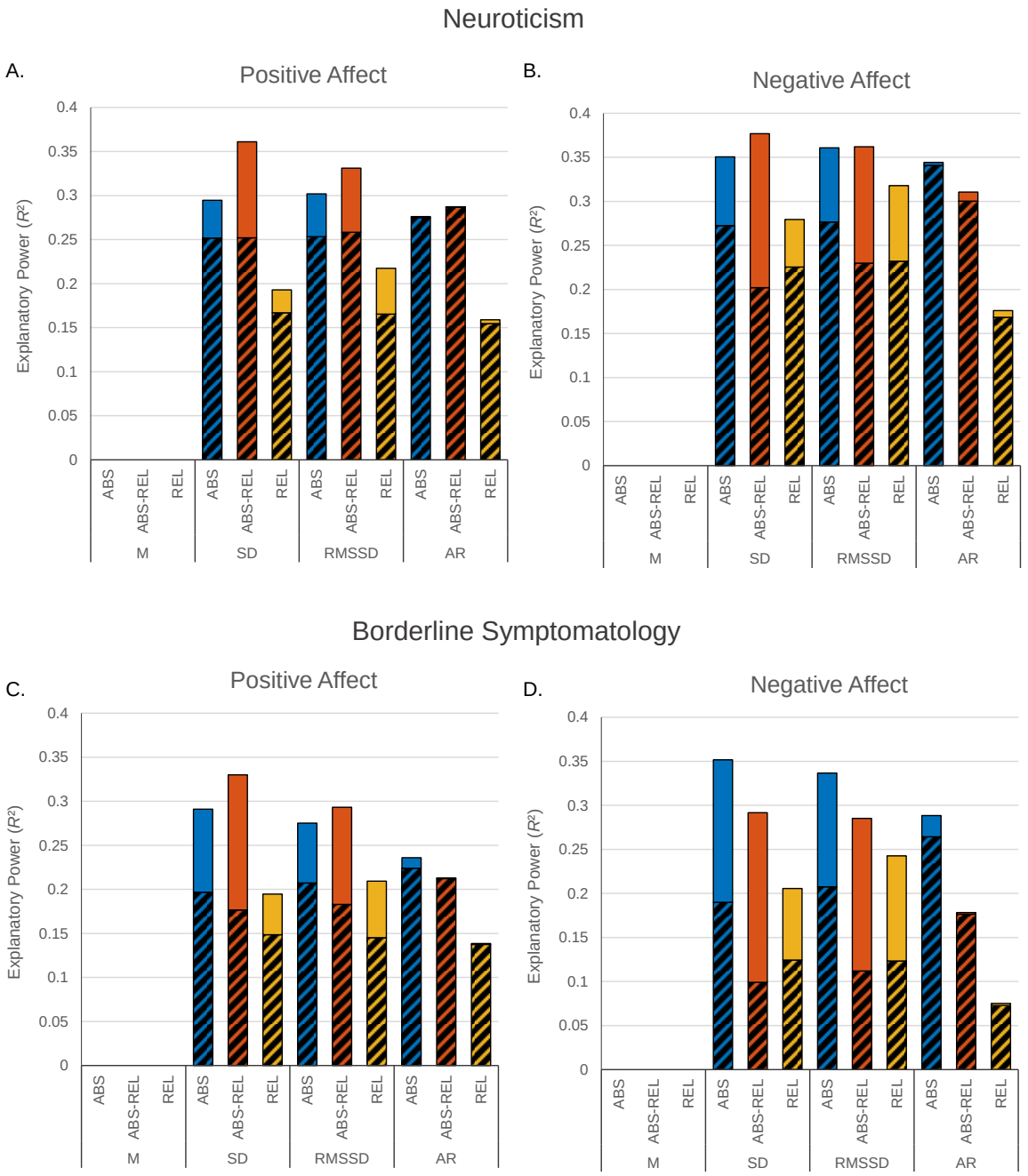


Figure 8. Evaluating the added external validity of people's emotion time series following each measurement condition. The two-by-two graph depicts the explanatory power (R^2) for each emotion dynamic measure (i.e., variability, instability and inertia) computed from a different time series type (ABS, ABS-REL, REL) in the prediction of people's neuroticism scores (top; n =176) and borderline symptomatology (bottom; n =171), for positive (left) and negative (right) affect, controlled for the predictive power of the mean level of PA or NA (which is computed from the time series of that measurement condition). The stacked bar chart indicates the total R^2 of a regression model, which comprises the relative contribution in R^2 of that particular emotion dynamic measure (plain) and the mean level of PA or NA of that measurement condition (shaded; see Supplementary Materials 10 for corresponding numerical values).