Affect dynamics or response bias?

The relationship between extreme response style and affect dynamics in a controlled experiment

Mirka Henninger[1], Niels Vanhasbroeck[2], & Francis Tuerlinckx[3]

[1] Faculty of Psychology, University of Basel

[2] Faculty of Social and Behavioural Sciences, Psychological Methods, University of Amsterdam

[3] Faculty of Psychology and Educational Sciences, Research Group of Quantitative Psychology and Individual Differences, KU Leuven

Author Note:

Funding: The data collection was performed by NV and FT and supported by the Research Fund of the KU Leuven (Grant C14/19/054) and by the Fonds Wetenschappelijk Onderzoek (FWO; Grant G074219N). The funders had no role in study design, data collection, analyses, decision to publish, or preparation of the manuscript. The authors declare that they have no conflict of interest.

Open Science and transparency: Data were originally published in Vanhasbroeck et al. (2024). Data collection has been preregistered (https://osf.io/ce87p/) and the data is publicly available (https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0123135/affective-consistency). Analysis scripts are publicly available (https://osf.io/w2kvp/?view_only=0b425909838d4b7cb5bb14c674360e5f).

Contact: Mirka Henninger, Missionsstrasse 62a, CH-4055 Basel; mirka.henninger@unibas.ch

Author contributions: Conceptualization: MH, NV, FT; Data curation: NV; Methodology: MH, NV, FT; Formal analysis: MH; Visualization: MH, NV, FT; Writing – original draft: MH; Writing – review and editing: MH, NV, FT.

The authors thank Timon Elmer for helpful comments on an earlier version of this manuscript.

Abstract

Intensive longitudinal data (ILD) have become a popular data format to capture people's momentary affect in everyday life. Besides describing persons' average affect over time, ILD are also often used to describe affect dynamics − that is how affect changes over time −, such as intraindividual variability or moment-to-moment temporal dependencies. Given that ILD studies mostly use self-report rating data, there is an increasing concern that response biases, such as extreme responding, might impact the estimates of affect dynamics. In this study, we assessed the relationship between extreme responding and affect dynamics in a controlled experiment. In a highly powered sample with $N = 1,398$ persons, we measured extreme responding using background questionnaires, and repeatedly induced affect using a probabilistic reward task with $T = 140$ trials per person. Our results suggest that people with high extreme response style trait levels show substantially higher measures of affect variability. However, extreme responding is neither associated with moment-to-moment temporal dependencies nor with participants' reactivity to affective stimuli. We conclude with a discussion on the importance of evaluating measurement in ILD for psychological assessments, and outline potential areas for future research to improve assessments of affect dynamics.

Keywords: Affect dynamics, intensive longitudinal data, measurement, response biases, extreme response style

Public Significance Statement: We investigate whether extreme response style is associated with affect dynamics in intensive longitudinal data. Our results show that extreme response style is associated with greater affect variability, but not with moment-to-moment temporal dependencies or reactivity to affective stimuli. Our study emphasizes the importance of evaluating measurement biases in psychological assessments.

Intensive longitudinal data (ILD) – data that consist of many observations within a same person across time – have become a popular data structure in psychological research. And with good reason, as these data allow researchers to uncover the dynamical processes that guide a person's behavior and mental states. Through understanding the dynamics of psychological processes, researchers are able to better understand how and why people develop psychopathological disorders such as depression or bipolar disorder, what characterizes these disorders, and how to describe and explain subclinical affective processes (e.g., McKone & Silk, 2022; Röcke & Brose, 2013; Russell et al., 2007).

One field in which ILD have shaped research is affect dynamics, which investigates how affective experiences change over time. These fluctuations are often thought to reflect how a person responds to external events and are considered important indicators of psychological functioning and well-being (Hedeker & Mermelstein, 2020; Koval et al., 2016; Kuppens et al., 2007; Röcke & Brose, 2013). Characteristics of affect dynamics, such as affect variability, are crucial in clinical research and have been associated with clinical outcomes, such as suicidal ideation (Palmier-Claus et al., 2012), mental health and borderline symptoms (Jenkins et al., 2024; Russell et al., 2007), or substance abuse (Mohr et al., 2015).

Within this domain, researchers typically make use of different types of measures to investigate affect dynamics (see e.g. Dejonckheere et al., 2019; Jahng et al., 2008; Wang et al., 2012; Wendt et al., 2020, for overviews). One often-investigated dynamical characteristic of affect is its intraindividual variability (also sometimes called dispersion), which is commonly operationalized as the within-person standard deviation (see e.g. Eid & Diener, 1999; Kuppens et al., 2007; Price et al., 2023). Intraindividual variability reflects how much a person's momentary affect deviates from their baseline affective

state (i.e., one's own average affect). In other words, it reflects the amplitude of fluctuations while ignoring the temporal dependencies of responses (e.g., Eid & Diener, 1999; Hedeker & Mermelstein, 2020; Kuppens et al., 2007; Larsen & Diener, 1987; Wichers et al., 2015).

Another measure of how affect changes over time is the first-order autoregressive effect, which captures temporal relations between adjacent time points. The first-order autoregressive effect, also called moment-to-moment predictability or inertia, indicates how strongly an affective state at a certain time point predicts the affective state at the following time point. It hence reflects how much a person's affective state persists from one moment to the next. In other words, it can be considered as the resistance to changes in affect over time, such that high levels of inertia might indicate a lack of affective flexibility (Bos et al., 2019; Hedeker & Mermelstein, 2020; Koval et al., 2013, 2016; Wang et al., 2012; Wichers et al., 2015).

Through the use of such measures of affect dynamics, researchers have been able to characterize people's affective lives. For example, research shows that both characteristics of affect dynamics reflect how a person adapts to their environment, and that the strength of affect dynamics remains relatively consistent over time (Bos et al., 2019; Carver, 2015; Kuppens & Verduyn, 2017; Panksepp, 2012; Smit et al., 2023). Additionally, research has associated increased affect variability with an increase in depressive symptoms, a higher probability of depression relapse, an overall lower well-being, lower emotional flexibility, and a lower self-esteem (Bos et al., 2019; Czyz et al., 2021; Franck & De Raedt, 2007; Hamaker et al., 2016; Hedeker, Mermelstein, et al., 2009; Houben et al., 2015; Koval et al., 2016; Maher et al., 2019; Piasecki et al., 2016; Trull et al., 2015). Thus, ILD have helped researchers gain new insights into the emotional system,

as well as into differences in how this system works across individuals (Ernst et al., 2021; Scott et al., 2020; Wang et al., 2012).

However, most of the discussed studies rely on self-report data to investigate measures of affect dynamics (e.g., using experience sampling methodology, ESM, or ecological momentary assessments, EMA). These measures might therefore be prone to influences by how participants use and interact with the measurement instrument, for instance, the rating scale. This means that these measures might not only capture affect dynamics, but might also be systematically impacted by a person's response styles (e.g., Adams et al., 2019; Baumgartner & Steenkamp, 2001; Bolt & Johnson, 2009). Examples of such response styles are the preference for extreme categories (i.e., extreme response style, ERS) or the middle category (i.e., midscale response style, MRS), and the tendency to agree with an item independent of the item content (i.e., acquiescent response style, ARS; e.g., Couch & Keniston, 1960; Hamilton, 1968; Henninger & Meiser, 2020a; Paulhus, 1991; Van Vaerenbergh & Thomas, 2013).

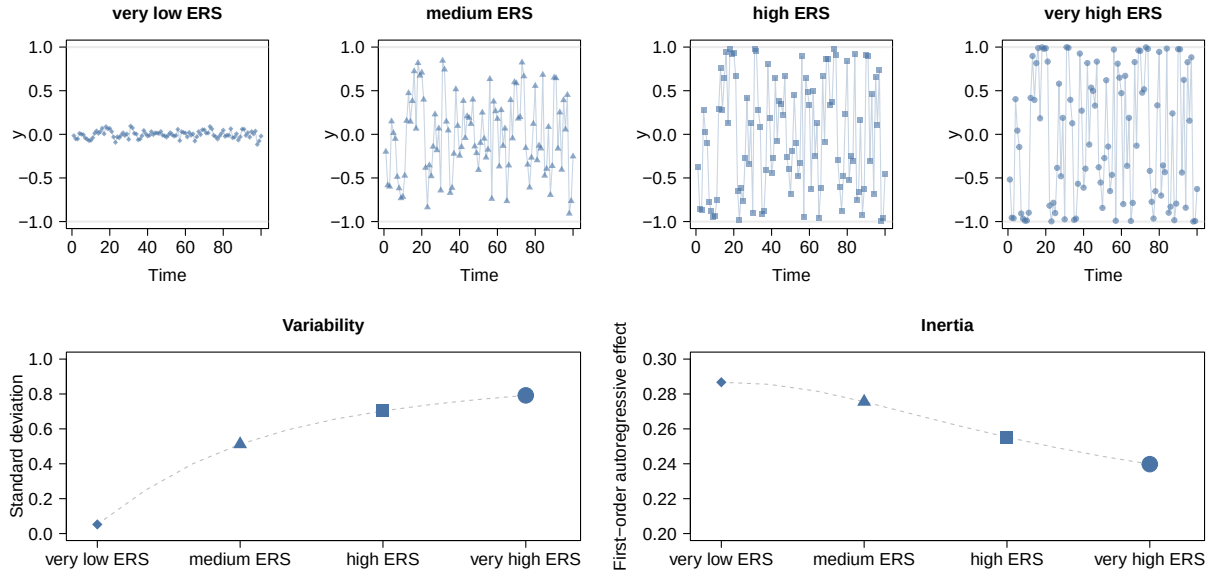
Among these response styles, ERS seems to be the most ubiquitous and prominent in self-report rating data and can have a substantial impact on the measures derived from this type of data. For instance, when response styles are examined in an exploratory manner, ERS tends to be omnipresent, while MRS and ARS are found to be less influential (e.g., Bolt et al., 2014; Falk & Cai, 2016; Henninger, 2021; Henninger & Meiser, 2020b, 2022; Plieninger & Heck, 2018). Furthermore, response styles have been shown to be stable over content domains and over time (Van Vaerenbergh & Thomas, 2013; Weijters et al., 2010a; Wetzel et al., 2013). ERS can therefore be considered as an additional variance component that is not related to the characteristic being measured, and carries the risk to bias estimates of trait scores, variability, and correlations (e.g.,

Adams et al., 2019; Böckenholt & Meiser, 2017; Bolt & Johnson, 2009; Bolt et al., 2014; De Jong et al., 2008; Plieninger, 2017). For these reasons, the awareness that controlling for the effects of response styles is crucial to safeguard data quality in self-report measures has increased in recent years.

In the context of ILD, response tendencies such as ERS might impact measures of variability. People with high ERS levels tend to overselect the extreme categories, and be thus assigned higher variability values − not because their affective states vary more strongly over time, but because of the way they use the rating scale. In contrast, people with low ERS levels tend to prefer the intermediate categories, hence provide responses at the center of the rating scale, which might result in reduced measured intraindividual variability. Unlike intraindividual variability, the relationship between ERS and inertia is more difficult to predict and may be less pronounced.

We use simulated, artificial data to illustrate potential effects of ERS on affect dynamics in Figure 1. In the top row, we show how affect measures in ILD may differ for different levels of ERS (assuming that the average person's affect location is 0). We can see that when ERS is low, the responses lie in the center of the scale (y-axis ranging from -1 to 1). When ERS is high, the full range of the scale is used and many responses are located at the extremes of the scale. The bottom row of Figure 1 shows the standard deviation (variability, left panel) and the first-order autoregressive effect (inertia, right panel) of the time series as a function of different ERS levels. In this simulated data, we observe that variability increases with higher ERS, while inertia decreases, albeit to a lesser extent. Please note that we simulated extreme cases (very low ERS, very high ERS) to lucidly illustrate *potential* effects of ERS on affect dynamics, and that in empirical data, the effects of ERS will, most likely, be less pronounced.

Figure 1: Illustration of how extreme responding (ERS) influences measures in a time series (top row) as well as variability and inertia (bottom row) based on a single artificial, simulated AR(1) time series. The observed variable y is considered to be continuous and bounded between -1 and 1. The same sequence is used to generate the four time series, but it is nonlinearly mapped to the [−1,1] interval using $\tanh(A * x)$, where A is the ERS level.



The relationships between ERS and different measures of affect dynamics in ILD have not been the focus of methodological studies yet – to the best of our knowledge, only two studies have examined the relationship between ERS and variability, and we have not encountered any studies addressing its relationship with inertia.

Firstly, Baird et al. (2017) asked participants to rate to which extent ten adjectives could describe cartoon characters from the Simpson television show. Because the cartoon characters are identical for each person, interindividual differences in variability are considered as response styles by the authors. Their results show that indeed the variability of the cartoon character ratings correlate moderately with measures of within-scale variability using personality self-reports. These results indicate that variability might be a person-specific response tendency that is stable across domains and might reflect components different from the measures of interest.

One limitation of this study is that it did not make use of a longitudinal data: Intraindividual variability was measured using the average standard deviation in personality self-reports across the Big Five traits or across different contexts (friends, family member, partner, student). In a second study, however, the authors did employ a daily diary design over a period of three weeks. The results of this study suggested that intraindividual variability of different constructs are associated with each other. Unfortunately, however, this study only included a limited number of measurement occasions ($T = 14$). Additionally, variability and response bias were both operationalizing through the standard deviation, which may potentially confound these measures and lead to an overestimation of the association between intraindividual variability and ERS. It would hence be desirable, to be able to clearly separate the measurement of affect dynamics and response biases, such as ERS.

Secondly, Deng et al. (2018) examined the relationship between ERS and intraindividual variability in positive and negative affect based on ten measurement occasions dispersed over an eight-week period. Using a psychometric modeling approach, the authors showed that the biasing effect of ERS on affect measures was highest for people with moderate values of positive and negative affect. The authors also highlighted the need to control for the biasing effects of ERS on affect variability using psychometric models.

A limitation of this study is that the measurements of affect were skewed, and thus showed a restriction of range in the observed intraindividual variability. As a consequence, measures of variability are likely to be associated with a person's mean value, potentially limiting the conclusions that can be drawn from the study (Bos et al., 2019; Koval et al., 2013; Mestdagh et al., 2018; Wagenmakers & Brown, 2007). Hence, it

would be important to separate affect location from intraindividual variability. In addition, both, ERS and affect variability, were measured using the same items (i.e., items measuring positive and negative affect), making it very likely that the ERS measure is confounded with the affect measures. High ERS naturally leads to higher proportions of choices in the extreme categories. When the distribution of affect is skewed (e.g., more people report high rather than low levels of positive affect), it becomes difficult to distinguish whether a response in the "strongly agree" category is driven by high ERS or high levels of affect. To fully separate the measurement of ERS from affect measures, it would be desirable to clearly disentangle affect location from variability, for instance by experimentally controlling affect location through affective stimuli and measuring ERS by a separate set of items.

The current study

In this study, we assess the relationship between ERS and different types of affect dynamics in a controlled experimental setting. By means of an experimental study, we can address the limitations in the studies by Baird et al. (2017) and Deng et al. (2018) mentioned above: (a) we obtain ERS trait scores from a psychometric model using items that are independent from the affect measures in the experiment, thereby avoiding a confound of ERS with affect measurement, (b) we use experimental stimuli of wins and losses to induce affect levels and disentangle location from variability, and (c) we aim at a high power using a large sample and a high number of measurement occasions to study multiple types of affect dynamics simultaneously. In the remainder of this section we discuss how we address each of these points in more detail.

First, to obtain valid ERS trait scores, we use item responses that are separate from

the affect measures collected during the experimental trials. This procedure follows the recommendations to use separate items in order to avoid confounding effects of trait and ERS measurement (e.g., Bolt et al., 2014; Henninger et al., 2023; Paulhus, 1991; Wetzel & Carstensen, 2017) and is based on the assumption that ERS is stable over content domains and time (Weijters et al., 2010b; Wetzel et al., 2016). Furthermore, we will apply state-of-the-art response style modeling using a so-called *IRTree* model (Böckenholt, 2012; De Boeck & Partchev, 2012; Jeon & De Boeck, 2019; Meiser et al., 2019). This approach allows us to obtain latent estimates of ERS levels that can be used as person-specific predictors for a person's intraindividual variability and inertia.

Second, instead of controlling for average affect levels when computing or assessing intraindividual variability scores (as e.g., Baird et al., 2017; Bos et al., 2019), we use data in which person levels of affect show intermediate values, thereby avoiding floor and ceiling effects. This property of the data has been achieved by directly inducing affect levels through experimental stimuli using a probabilistic reward task (Vanhasbroeck et al., 2024, see also Koval et al., 2016, for another laboratory approach to induce affect). In this task, participants are exposed to wins and losses as outcomes in each experimental trial. Affect ratings are collected after the amount of win or loss has been revealed to the participant. This experimental induction of affect solves the confound between intraindividual variability and average affect location. Because the obtained distribution of affect measures is approximately uniform, we avoid that persons show low or high affect locations resulting in floor or ceiling effects and reduced variability (see Baird et al., 2017; Deng et al., 2018). Thus, using controlled experimental stimuli for affect, we are able to observe affect ratings over the whole response range for all participants.

Third, modeling affect dynamics requires a larger sample size compared to solely modeling affect location (see Wang et al., 2012, for a discussion). In this study, we used data from $N = 1,398$ persons and $T = 140$ trials. This large sample size with a high number of measurement occasions allows us to employ a latent variable modeling approach to study intraindividual variability and inertia simultaneously (Hedeker & Mermelstein, 2020; Wang et al., 2012), and to include additional random effects to study the relationship between ERS and reactivity to affective stimuli.

Research questions

Our main interest lies in exploring whether persons with high ERS levels exhibit higher intraindividual variability in affect measures after controlling for the previous affect rating and the stimulus strength (RQ1). To study this question, we will assess the extent to which the stable person-specific tendency to give extreme responses is associated with intraindividual variability in affect. If it is the case that some of the variability in affect measurement in ILD is due to a general tendency to provide extreme responses, this result might suggest that researchers who associated affect variability measures with, for example, psychopathological development, tap into explaining ERS rather than affect variability. Thus, implications based on research findings with regard to affect variability might be limited when ERS is not accounted for.

Second, we will examine the relationship of ERS and inertia (RQ2), which, to our knowledge, has not yet been examined in previous studies. Intraindividual variability and inertia capture two different aspects of affect dynamics with low intercorrelations (Wang et al., 2012). We will explore whether and how ERS is associated to inertia (i.e. the retraction to the average affect level) operationalized by the first-order autoregres-

sive parameter.

Third, we will explore whether ERS levels are associated with participants' reactivity to affective stimuli. A preference for extreme categories has been shown to be associated with intensified judgments (Böckenholt, 2012; He & van de Vijver, 2016; Mourali et al., 2007; Weijters et al., 2016). Other studies have suggested relationships of ERS and personality or cultural characteristics, suggesting that different ERS levels may be due to differential perception of the environment (He & Van De Vijver, 2013; Hoffmann et al., 2013; Plieninger, 2020; Ulitzsch et al., 2024). We will explore whether people with high ERS show higher reactivity to affective stimuli (i.e. the outcome, namely wins and losses in the probabilistic rewards task). Such a finding could suggest that people with high ERS do not merely use the rating scale differently but may actually react distinctively to affective stimuli.

Methods

Transparency and Openness

This study relies on publicly available experimental data that were collected by Vanhasbroeck et al. (2024). In their study, the authors investigated the reliability of several single-item measures of affect by means of an online experiment. We summarize the key aspects of the experimental study here and refer to the original paper for a full description of the data and methods. The experimental design, tasks, procedures, and measures have been preregistered,[1] and the data is publicly available.[2]

For the current study, the research questions and data analysis scripts have not been

[1] https://osf.io/ce87p/
[2] https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0123135/affective-consistency

preregistered. Mplus output files, Markov Chain Monte Carlo (MCMC) traceplots, and results of the robustness checks are available on OSF.[3]

Sample characteristics and procedure

A total of 1,398 participants were recruited through the online data collection platform Prolific (https://www.prolific.co/). 60% of the participants identified as male and 68% enjoyed a higher education. Participants were on average 34 years old ($SD = 13$, range $= [18, 76]$). Participants were told that they would receive their total task earnings as a reward for their participation. Unbeknownst to them, this total was predetermined and set to £5. Participants spent on average 23min 59s (SD = 9min 6s, range = [11min 21s, 79min 39s]) on the task.

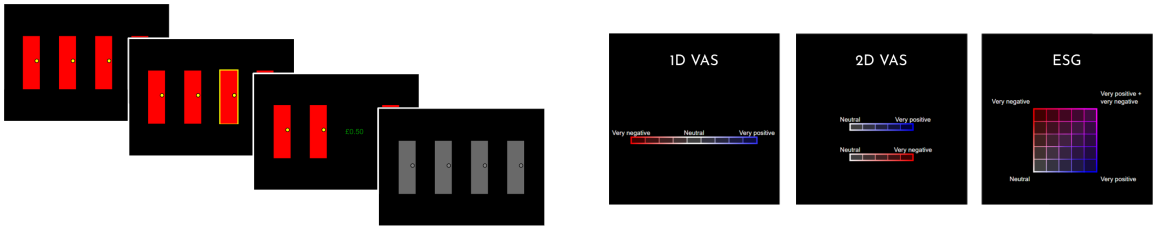
After giving consent, participants completed a self-paced probabilistic reward task (see Vanhasbroeck et al., 2024). In each of the $T = 140$ trials, they were presented with four doors (see Panel A in Figure 2). They were told that two of these doors hid a monetary reward while the other two doors hid a monetary loss. On each trial, participants had to choose one of the doors, then the door lit up for 500ms and finally opened, displaying the monetary outcome (between $-£0.5$ and $£0.5$) for 2s. While participants were told that they had an equal probability of winning on each trial, the monetary outcomes were in reality predetermined, though the order in which participants received them varied between participants.

Next, participants were asked to report their affective state using a rating scale of a specific format (see Panel B in Figure 2 and below). Participants were asked to indicate the "position on the scale [which] provides information on how positive (blue)

[3]https://osf.io/w2kvp/?view_only=0b425909838d4b7cb5bb14c674360e5f

Figure 2: Illustration of the experimental procedure. Panel (A) shows the sequence in each trial: (1) choosing a door, (2) the choice is being displayed through a yellow frame around the door, (3) the door opens and the outcome is displayed, (4) the doors become inactive until the participant has rated their affective state. Panel (B) shows the different rating scale formats that were used to measure affect.

(A) **(B)**



Note: 1D VAS: one-dimensional visual analogue scale; 2D VAS: two-dimensional visual analogue scale; ESG: evaluative space grid. Note that rating scale formats were between individuals in the original study, but this variation was not considered in this study.

and/or negative (red)" a participant feels (exact wording in quotes). Responses of the participants remained visible on screen as a red dot. Participants could move on to the next trial by pressing the spacebar. The red dot disappeared at the start of the next trial.

At the end of the study, participants were asked to fill out two questionnaires, namely a questionnaire about how they experienced the study that we used to estimate an ERS score for each person, as outlined in detail further below, and a depression questionnaire that is not used in our analyses and, hence, is not discussed any further. After filling out both questionnaires, participants were debriefed and given a diligence question ("In your honest opinion, should we use your data?"), to which they could respond with "yes" or "no". People who answered "no" are not included in this dataset (see Vanhasbroeck et al., 2024, for more details).

Materials

Measurement of affect

Vanhasbroeck et al. (2024) were interested in the reliability with which participants report on their momentary affective states. For this reason, the authors varied the visual layout of the rating scale (referred to as a format) and whether it could either be answered in a continuous or discrete manner (referred to as continuity) randomly between-participants (see Panel B in Figure 2 and Vanhasbroeck et al., 2024, for more details). Since Vanhasbroeck et al. (2024) did not find any indication that these response formats influenced measurement consistency (see also Henninger et al., 2023, for similar results for the effect of response formats on response scale use), we did not further consider the variation between the response formats in our study (apart from robustness checks.

Note that positive and negative affect (PA/NA) were measured with two types of scales (ESG and 2D VAS), while valence was directly measured using one scale (1D VAS), and indirectly measured using the other two scales.[4] In line with Vanhasbroeck et al. (2024) and to be able to consider the responses on all scales together, we calculated a one-dimensional valence score for the ESG and 2D VAS response formats as:

$$\text{valence}_t = 0.5 * (\text{PA}_t - \text{NA}_t + 1), \tag{1}$$

where the values of valence are high when PA is high and NA is low and vice versa. For instance, when PA = 1 and NA = 0, valence = 1, when PA = 0 and NA = 1, valence = 0, and when PA = NA, valence = $\frac{1}{2}$. Note that this derived measure also

[4]Note that all measurements that were obtained through these measures fall in the range from 0 to 1.

falls within the range of 0 and 1, and that this calculation implies that measurements from the ESG and 2D VAS are included in our analyses for positive and negative affect, and additionally for the derived valence scores.

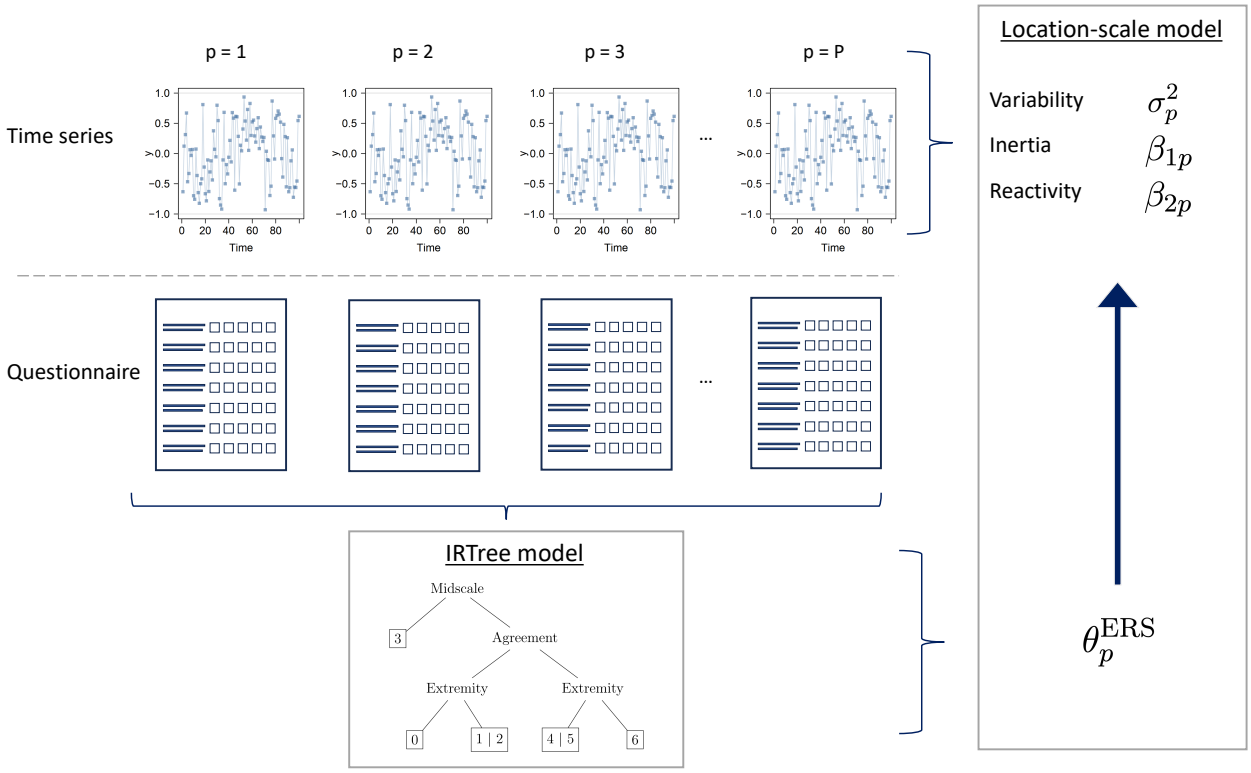
Questionnaire used to measure ERS

To estimate ERS trait scores, we use measurements of how participants experienced the experiment assessed by a separate set of items. Vanhasbroeck et al. (2024) asked participants to fill out a self-designed, unvalidated questionnaire at the end of the study. The questions targeted the extent to which participants were influenced by the stimuli of the experiment (Q1-Q4), how motivated participants were to complete the study (Q5-Q9), and the ease and accuracy with which participants felt they could report their feelings (Q10-Q17). Participants had to answer each of these questions on a 7-point Likert scale from "strongly disagree" to "strongly agree". The questions were presented in the same order to all participants and are presented in Table 4 in Appendix A.

Analysis approach

We employ a joint modeling approach to assess the influence of ERS on affect dynamics. Herein, we modeled the latent ERS trait using a psychometric modeling approach for response styles, namely a two-parameter IRTree model (Böckenholt, 2012). Within the same modeling approach, we use this latent ERS trait estimates as a predictor in a multilevel location-scale regression model to explain interindividual levels of intraindividual variability, of inertia, and of the influence of the experimental outcomes on affect (Hedeker, Demirtas, & Mermelstein, 2009; Hedeker & Mermelstein, 2020; McNeish, 2021). Figure 3 illustrates our analysis approach graphically.

Figure 3: Illustration of our analysis approach: ERS trait scores θ_p^{ERS} are estimated from questionnaire responses using an IRTree model. They serve as a predictor in a location-scale model to explain affect dynamics.

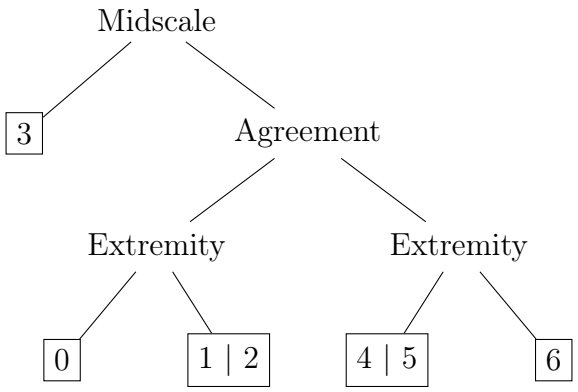


Psychometric model for ERS (IRTree)

Response styles can be modeled using extensions to modeling approaches from item response theory that incorporate them as additional latent dimensions (see Böckenholt & Meiser, 2017; Henninger & Meiser, 2020a, 2022, for reviews). A popular family of response style models are IRTree models. They decompose the responses to rating scale items into distinct response processes (Böckenholt, 2012; De Boeck & Partchev, 2012; Khorramdel & von Davier, 2014; Meiser et al., 2019; Plieninger & Meiser, 2014). For a 7-point item with response categories 0 to 6, the responses can be decomposed into a midscale response process (MRS; responses in the middle category 3 vs. non-mid responses in categories 0 - 2 and 4 - 6), an agreement response process (responses in

the disagreement categories 0 - 2 vs. the agreement categories 4 - 6), and an extremity response process (ERS; responses in the extreme categories, 0 and 6 vs. the intermediate categories 1, 2, 4, 5). Figure 4 illustrates the decomposition of the observed response on a 7-point Likert-type scale into the midscale, agreement, and extremity response processes using an IRTree.

Figure 4: IRTree structure for 7-point rating items with midscale, agreement and extremity response processes.



As a preparatory step to estimate the IRTree model, the rating data (in our case, the rating data of the background questionnaire) is decomposed into pseudo-items reflecting the response processes, as presented in the left and middle part of Table 1. For the midscale response process, all disagreement and agreement categories are coded as 0, while the middle response category 3 are coded as 1. Similarly, for the agreement response processes, the disagreement categories $0-2$ are coded as 0, while the agreement categories $4-6$ are coded as 1. For the extremity response processes, the intermediate categories 1, 2, 4, 5 are coded as 0, and the extreme categories 0 and 6 are coded as 1. As we will outline further below, we will use the factor scores of the extremity response process as the ERS predictor variable in the location-scale model.

Table 1: Coding of pseudo-items for a 7-point Likert-type scale for the midscale, agreement, and extremity response processes and parameterization using the two-parameter logistic model. The observed response of persons p on item i is denoted by X_{pi}

	X_{pi}							
	0	1	2	3	4	5	6	
Midscale (Y_{1pi})	0	0	0	1	0	0	0	$\frac{\exp(\alpha_{1i}\theta_{1p}-\delta_{1i})}{1+\exp(\alpha_{1i}\theta_{1p}-\delta_{1i})}$
Agreement (Y_{2pi})	0	0	0	–	1	1	1	$\frac{\exp(\alpha_{2i}\theta_{2p}-\delta_{2i})}{1+\exp(\alpha_{2i}\theta_{2p}-\delta_{2i})}$
Extremity (Y_{3pi})	1	0	0	–	0	0	1	$\frac{\exp(\alpha_{3i}\theta_{3p}-\delta_{3i})}{1+\exp(\alpha_{3i}\theta_{3p}-\delta_{3i})}$

The right part of Table 1 shows how these three sets of pseudo-items, each reflecting a specific response process, are parameterized using a two-parameter logistic model from item response theory. The model includes separate trait parameters for persons (θ_p) and parameters reflecting item difficulty (δ_i) and discrimination (α_i) for each of the three response processes. Hence, persons' trait parameters for the midscale (θ_{1p}), agreement (θ_{2p}), and extremity (θ_{3p}) response process are estimated. It also means that the pseudo-items that describe each process can be differentially difficult (e.g., δ_{1i} reflects how much item i fosters midscale responses, while δ_{3i} reflects how much item i fosters extreme responses). Items can also have a differential impact on the latent traits through discrimination parameters α_i.

The probability of observing a response in category k, here with $k \in 0, \ldots, 6$, is given from multiplying over the model probabilities of all response processes, like following down the branches of the IRTree presented in Figure 4 (see Böckenholt, 2012; Plieninger,

2020):[5]

$$
\begin{aligned}
p(X_{pi} = k) = & \frac{\exp(y_{1pi}(\alpha_{1i}\theta_{3p} - \delta_{1i}))}{1 + \exp(\alpha_{1i}\theta_{1p} - \delta_{1i})} \\
& \times \left(\frac{\exp(y_{2pi}(\alpha_{2i}\theta_{2p} - \delta_{2i}))}{1 + \exp(\alpha_{2i}\theta_{2p} - \delta_{2i})} \right)^{(1-y_{1pi})} \\
& \times \left(\frac{\exp(y_{3pi}(\alpha_{3i}\theta_{3p} - \delta_{3i}))}{1 + \exp(\alpha_{3i}\theta_{3p} - \delta_{3i})} \right)^{(1-y_{1pi})}
\end{aligned}
$$

In the specified three-process IRTree model, we fixed the means of the latent dimensions to 0 and estimated the full variance-covariance matrix, hence allowing midscale, agreement, and extremity processes to be associated with each other.

We used the IRTree modeling approach to obtain a latent estimate of the ERS trait scores. These factor scores of the extremity response process (θ_{3p}) serve as the ERS predictor variable θ_p^{ERS} in the location-scale model. For this purpose, we created pseudo-items using the self-reported measures based on the questionnaire about participants' experience with the experiment. Note that we conducted different types of robustness checks for the IRTree model that are presented in Appendix B and the supplementary materials on OSF.

Location-scale model using ERS as a predictor variable

Our main analysis model is a multilevel location-scale model for longitudinal data (Hedeker & Mermelstein, 2020; McNeish, 2021; Wang et al., 2012). This model allows us to model not only interindividual differences in mean levels of affect, but also their intraindividual variability (RQ1), inertia (RQ2), and reactivity to affective stimuli

[5]Note that the second and third process are taken to the power of $(1 - y_{1pi})$. This characteristic ensures that in case the response is in the middle category (first response process), the second and third response processes are not evaluated, and vice versa (see also Table 1).

(RQ3). The model is specified on two levels, where Level-1 represents the experimental trials and Level-2 represents the participants of the experiment.

On Level-1, we explain affect measures (y_{tp}) for trial t and person p by specifying an autoregressive effect (inertia, β_{1p}) and an effect of participants' reactivity to the affective stimuli (i.e. the amount of win or loss in each experimental trial, called outcome, β_{2p}):

$$\text{Level-1:} \qquad y_{tp} = \beta_{0p} + \beta_{1p} \cdot y_{(t-1)p} + \beta_{2p} \cdot \text{outcome}_{tp} + \varepsilon_{tp}$$

We allow the Level-1 variance to be specific for each person p to obtain a measure of intraindividual variability:

$$\varepsilon_{tp} \sim N(0, \sigma_p^2)$$

We explain between-person differences in intraindividual variability using the latent ERS trait variable as a predictor:[6]

$$\text{Level-2 model for variability/scale (RQ1):} \qquad \sigma_p^2 = \exp(\omega_0 + \omega_1 \cdot \theta_p^{\text{ERS}} + u_{3p}) \qquad (2)$$

On Level-2, we specified random effects for the intercept (β_{0p}), the slope of the autoregressive effect (β_{1p}), and the slope of the outcome variable (β_{3p}). This means that the locations of the affect variables, inertia, and reactivity to affective stimuli are allowed to vary between persons. We used the latent ERS trait variable (θ_{3p}) estimated in the IRTree model as a predictor variable for the between-person intercept, and the

[6]Note that the exponential function ensures that the variance σ_p^2 is positive and that when u_{3p} follows a normal distribution, σ_p^2 is lognormally distributed at the level of the person (Hedeker, Demirtas, & Mermelstein, 2009; Hedeker et al., 2008).

between-person slopes of inertia and outcome:

Level-2 model for location:

$$\begin{aligned}
&\text{Explain random intercept:} && \beta_{0p} = \gamma_{00} + \gamma_{01} \cdot \theta_p^{\text{ERS}} + u_{0p} \\
&\text{Explain random slope of inertia (RQ2):} && \beta_{1p} = \gamma_{10} + \gamma_{11} \cdot \theta_p^{\text{ERS}} + u_{1p} \\
&\text{Explain random slope of outcome (RQ3):} && \beta_{2p} = \gamma_{20} + \gamma_{21} \cdot \theta_p^{\text{ERS}} + u_{2p}
\end{aligned}$$

The variance components (random intercept and slopes, and intraindividual variability) follow a multivariate normal distribution with means fixed to 0 and a variance-covariance matrix Σ_τ:

$$\begin{bmatrix} u_{0p} \\ u_{1p} \\ u_{2p} \\ u_{3p} \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & & & \\ \tau_{01} & \tau_{11} & & \\ \tau_{02} & \tau_{12} & \tau_{22} & \\ \tau_{03} & \tau_{13} & \tau_{23} & \tau_{33} \end{bmatrix} \right). \tag{3}$$

We estimate the latent ERS trait variable and affect dynamics in a joint modeling approach. Such a one-step approach is advantageous because it takes uncertainty in the estimation of latent person parameters into account when regressing ERS trait scores on intraindividual variability, inertia, and reactivity to affective stimuli (Hedeker, Demirtas, & Mermelstein, 2009; Hedeker & Mermelstein, 2020; Wang et al., 2012). Due to the complexity of the modeling approach with four random components and their covariance matrix, we estimated a separate models for positive affect, negative affect, and valence. Importantly, we do not account for the characteristics of the rating scales in our analyses, as they were shown to have no effect on the data characteristics (see

Vanhasbroeck et al., 2024). However, we conducted robustness checks to assess the impact of the visual layout of the rating scales for our main analyses showing that the results are largely unaffected (see Appendix B).

Model estimation

We estimated the models using Mplus 8.6 (L. K. Muthén & Muthén, 2017) with Bayesian estimation using the default uninformative prior distributions of Mplus in combination with prior sensitivity analyses (see Asparouhov & Muthen, 2023; Gelman & Hill, 2006; Guo et al., 2019; Leckie et al., 2014; Lin et al., 2018; B. Muthén & Asparouhov, 2012; Rast et al., 2012).[7]

In the IRTree model, we used normal prior distributions with mean 0 and standard deviation 5 for discrimination and difficulty parameters α_i and δ_i and inverse-Wishart priors for variances and covariances, i.e., $IW(1, 4)$ and $IW(0, 4)$, respectively. In the location-scale model, we used uninformative, univariate normal prior distributions with mean 0 and standard deviation 10^{10} for the regression parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$. For variances and covariances of random intercepts and slopes in the location-scale model, we used an improper inverse-Wishart prior $IW(0, -5)$ (Asparouhov & Muthen, 2010; Asparouhov & Muthen, 2023). We conducted prior sensitivity analyses for the inverse-Wishart priors and observed consistent parameter estimates across prior specifications, indicating the robustness of model estimates against prior specifications (see Appendix B and the supplementary material on OSF for more details).

The final models were fit using twelve chains each with dispersed starting values

[7]We furthermore used the R packages `MplusAutomation` (Hallquist & Wiley, 2018) for integrating model results into R, `mirt` (Chalmers, 2012) for marginal maximum likelihood estimation of the IRTree model to derive data-dependent priors for prior sensitivity analysis, and `ggplot2` (Wickham, 2016) for plotting.

using 5,000 iterations, with half of them as burn-in (Muthen, 2010). We assessed model convergence by inspecting overlaid Bayesian traceplots of the multiple chains (see supplementary material on OSF) and the potential scale reduction which was $PSR = 1.009$ in the model for positive affect, $PSR = 1.009$ in the model for negative affect, and $PSR = 1.017$ in the model for valence, indicating that the chains mixed well (Brooks & Gelman, 1998; Gelman & Rubin, 1992).

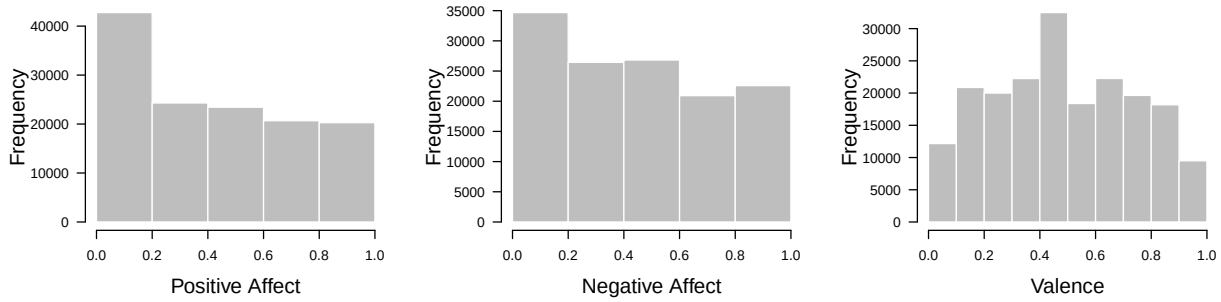
Results

Descriptive distributions of positive affect, negative affect, and valence

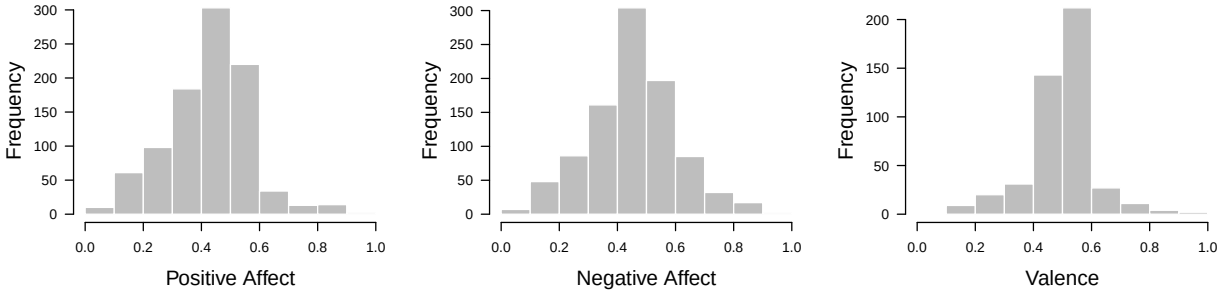
Figure 5 shows the distribution of affect measures on Level-1 in the top row and the distribution of affect person means in the bottom row. As expected, the full range of the response scale has been used by the participants in the experiment when providing responses on the affect measures (top row). Furthermore, the distribution of person means (bottom row) are peaked and symmetric around the center of the scale. This distribution reflects the fact that affect was manipulated through the probabilistic reward task in a balanced manner, effectively reducing the variability of person-specific affective means. However, the data shows substantial variability for Level-1 measures of affect, as one can see in the top row of this figure. Both characteristics of the affect distributions – lower variability in the means of affect and high intraindividual variability – are advantageous for the current analysis as discussed above.

Figure 5: In the top row, the distribution of all affect measures at the individual level is shown. In the bottom row, the distributions of the person means are displayed. The bins of the histograms in the top row are chosen in accordance to the response format (see Figure 2). Positive and negative affect were measured using five response categories, and are thus presented in five bins. The valence measure, in contrast, is based on a computed score (see Equation 1) resulting in nine bins.

Distribution of Level-1 affect measures



Distribution of person means



Results of the IRTree model used to estimate ERS

Table 2 shows the variances and correlations of latent traits in the IRTree model based on the items in the background questionnaire. The estimated parameters are highly consistent across the three estimated models (positive and negative affect, valence), and in line with previous studies (e.g., Henninger, 2021; Meiser et al., 2019; Plieninger & Heck, 2018).[8] We can see that the variance of the extremity response process is quite

[8]Note that due to our one-step appraoch, the IRTree model has been estimated together with the location-scale model. Hence, we obtain parameter estimate for the IRTree model for each affective variable (positive affect, negative affect, and valence), even though the IRTree model is based on the different set of items from the background questionnaire.

substantial, while variance of the midscale response process is comparably small. As expected, we see a negative correlation between the extremity and the midscale response processes, and small to moderate correlations between the extremity and the agreement response process. The latter indicates that extremity (i.e., ERS) is not capturing very high or low agreement, but is a distinct trait. The full list of estimated parameter with item and discrimination parameters can be accessed in the supplementary material on OSF.

Table 2: Variances (diagonals) and correlations (off-diagonals) of latent traits in the IRTree model estimated in a one-step approach with the location-scale model of positive affect, negative affect, and valence.

	Positive Affect			Negative Affect			Valence		
	Mid	Agree	Extremity	Mid	Agree	ERS	Mid	Agree	Extremity
Mid	0.10			0.10			0.10		
Agree	-0.55	0.42		-0.55	0.42		-0.56	0.44	
Extremity	-0.56	0.31	0.40	-0.56	0.31	0.41	-0.55	0.35	0.34

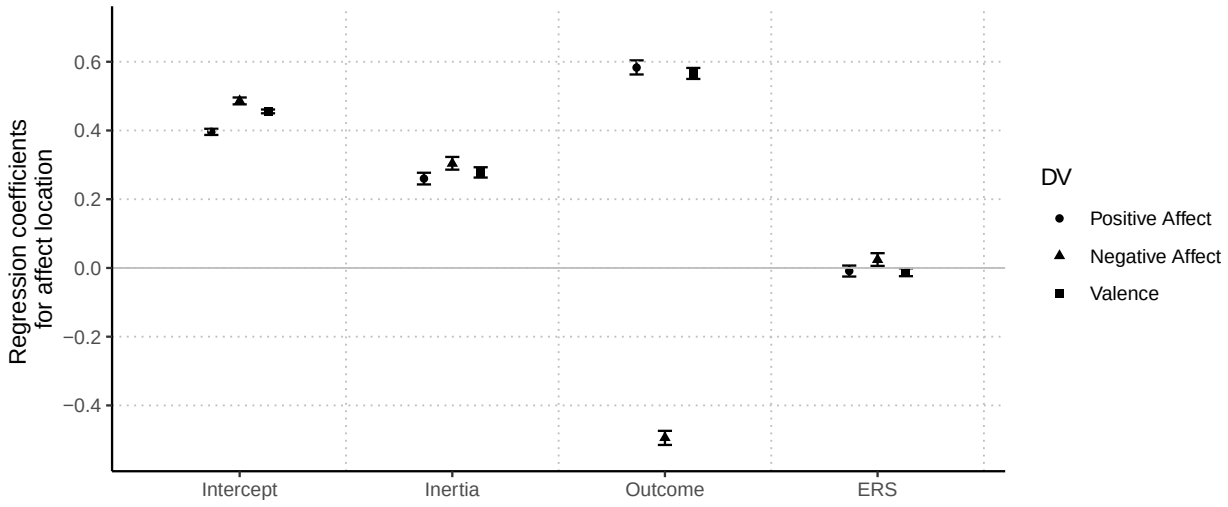
Note: Mid: Midscale response process; Agree: Agreement response process; Extremity: Extremity response process (ERS). Please note that the background questionnaire, which serves as the data foundation for the IRTree model, is independent of the affect measures. However, separate models are estimated for positive affect, negative affect, and valence, within which the IRTree model is also estimated using a one-step approach.

Fixed effects on location in the location-scale model

Figure 6 shows the fixed effects on a participant's locations of affective states throughout the experiment. The results across the three outcomes are very consistent and could be precisely estimated (small 95% credible intervals) because of the high power in the current study. The interested reader is also referred to Appendix C for the exact parameter estimates together with their 95% credible interval.

The results for the intercept, inertia, and outcome are presented mainly for plausibility checking and to ensure comparability to previous studies. The fixed intercepts

Figure 6: Fixed effect regression coefficients for affect locations.



Note: Error bars reflect 95% credible intervals (see Appendix C for exact values). DV: Dependent variable; ERS: Extreme response style.

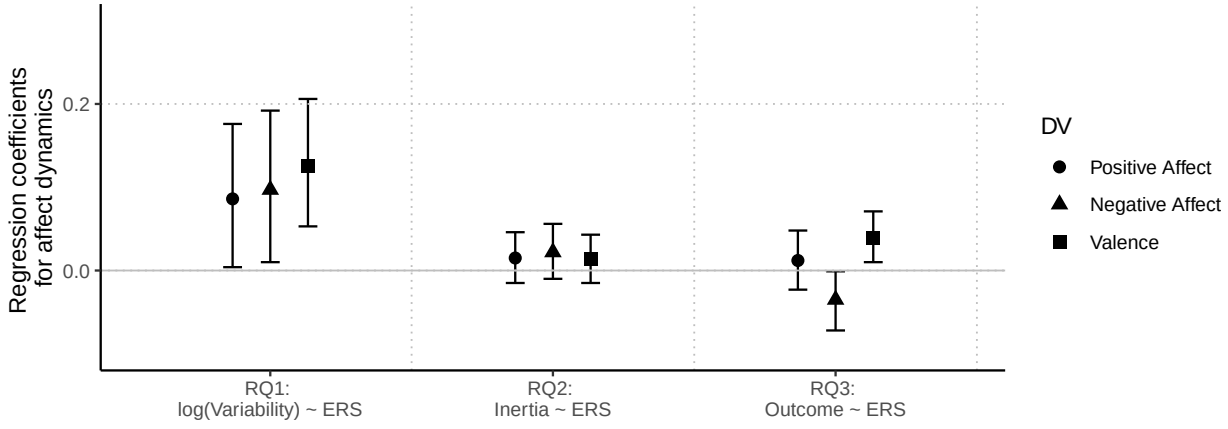
for the locations of all three affective variables lie around .5 which is plausible given the affect scale that ranged from 0 to 1 and we induced affect in a balanced manner. The fixed effect of inertia lies around .3 for all three affective variables, which is a typical estimate for inertia in affect research (e.g., Koval et al., 2016). The predictor variable outcome (i.e., amount of win or loss in each trial; reactivity) was associated with the affective variables in the expected direction (positively for positive affect and valence, negatively for negative affect).

Interindividual differences in the levels of ERS have no effect on interindividual differences in the location of affect. This result was to be expected, since we manipulated average affect levels in the experiment resulting in a reduced variability in person means of affect (see Figure 5).

Effect of ERS on affect dynamics in the location-scale model

Figure 7 summarizes the effect of ERS on the affect dynamical measures. First and foremost, we can see that ERS has a positive effect on log affect variability (RQ1). Respondents with higher ERS levels as measured by the background questionnaire tend to show larger within-person variances of affect across trials (Positive affect: $\hat{\omega}_1 = 0.09$, CI = [0.004, 0.18]; Negative affect: $\hat{\omega}_1 = 0.10$, CI = [0.01, 0.19]; Valence: $\hat{\omega}_1 = 0.13$, CI = [0.05, 0.21]). Regarding the second and third research questions, ERS does not seem to be consistently associated with between-person differences in inertia (RQ2) neither with reactivity to affect stimuli as reflected in the effect of ERS on the slope of outcome (RQ3).

Figure 7: Fixed effect regression coefficients for affect dynamics: Effect of ERS on variability (RQ1), inertia (RQ2), and outcome (i.e., reactivity to affective stimuli; RQ3).



Note: Error bars reflect 95% credible intervals (see Appendix C for exact values). DV: Dependent variable; RQ: Research Question; ERS: Extreme response style.

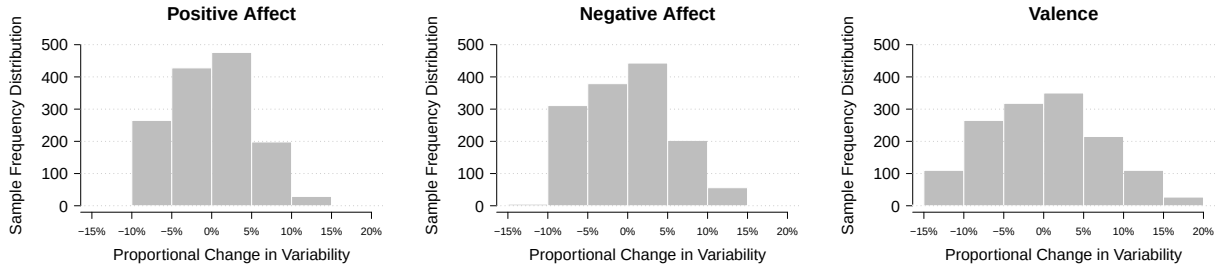
Proportional change in affect variability as a function of ERS

Because variability is estimated on the log scale (see Equation 2), its interpretation is more challenging. In this case, the exponential of the regression coefficient of ERS on

variability reflects the proportional change in the dependent variable and can serve as an effect size for the relationship between affect variability and ERS (see e.g., Hedeker et al., 2008; Lo & Andrews, 2015). This means that $\sqrt{\exp(\hat{\omega}_1)}$ is the proportional change that we expect in the standard deviation of affect for a one-unit increase in ERS. Hence, if ERS increases by 1 (which is a plausible change given that the estimated ERS levels range from -1.2 to 1.56 in our data), the expected proportional change in the standard deviation of affect amounts to 8.98% for positive affect, 10.19% for negative affect, and 13.43% for valence.

Figure 8 further quantifies the magnitude of the association between ERS and affect variability (as specified in Equation 2) in the present sample. The x-axis shows the model-predicted proportional change in affect variability for one person given their ERS trait score, the y-axis shows the frequency of occurrence of this effect size in the analysis sample. For instance, for valence, we can see that approximately 100 participants show a reduced variability by -10% to -15% due to their low ERS levels, while approximately 110 participants show an increased variability by 10% to 15% due to their high ERS levels in the analyzed dataset.

Figure 8: Sample frequency distribution of the model-predicted proportional change in affect variability.



Affect response distributions for different ERS levels

We further descriptively illustrate the relationship between ERS and the self-reported ratings (i.e., raw data) of positive affect, negative affect, and valence in Figure 9. We show the frequency distributions of the three affective variables as a function of ERS levels. The columns in the figure show these frequency distributions for participants with low (lowest 20%), moderate (middle 60%), and high (highest 20%) in the left, middle, and right panel, respectively. For all three affect variables, the proportion of responses in intermediate bins decreases and the proportion of responses in the extreme bins increases from the left over the middle to the right panel. In the right panel depicting affect measures for the 20% of the sample with the highest ERS level, most affect responses are given in the extreme bins at the lower and upper end of the scale.

Random effects in the location-scale model

Lastly, in Table 3 we present the variances and covariances between the random effects in the location-scale model as specified in Equation 3. As expected from a controlled experiment with balanced stimuli, the intercepts of affect show essentially no variation across individuals. Conversely, interindividual differences in variability is very pronounced, followed by interindividual differences in inertia and the effect of outcome.

Table 3 furthermore illustrates associations between the different measures of affect dynamics. First, residual variability is negatively correlated with residual variability in inertia. This means that people with higher affective variability tended to take longer to return to their baselines. Second, residual variability is associated with more positive residual slopes for positive affect and valence, and more negative residual slopes for negative affect. Third, participants with higher residual autoregressive effects showed

Table 3: Residual variances (diagonals) and correlations Σ_τ of random effects (off-diagonals) for negative and positive affect and valence (see Equation 3).

	Intercept	Inertia	Outcome	Variability
Positive Affect				
Intercept	0.02			
Inertia	-0.20	0.10		
Outcome	0.31	-0.70	0.07	
Variability	0.15	-0.37	0.66	0.54
Negative Affect				
Intercept	0.02			
Inertia	0.23	0.10		
Outcome	-0.11	0.61	0.08	
Variability	0.17	-0.38	-0.45	0.61
Valence				
Intercept	0.01			
Inertia	-0.28	0.09		
Outcome	0.08	-0.68	0.08	
Variability	-0.12	-0.35	0.50	0.55

Note: Credible intervals for variance and correlation estimates are presented in Appendix C.

less pronounced residual effects of outcome (less positive for positive affect and valence, less negative for negative affect). This correlation indicates that participants with higher inertia were less reactive to the monetary outcome of the current trial.

Discussion

In this study, we examined the association between ERS and different measures of affect dynamics in a controlled experiment. Using a location-scale model, our results suggest that ERS levels are positively associated with affect variability (RQ1). However, ERS is not associated with interindividual differences in inertia (RQ2), nor with participants' reactivity to the affective stimuli they encountered during the study (RQ3). Our study is the first highly powered study assessing the influence of response biases in ILD using unconfounded measurement of ERS and various measures of affect dynamics

simultaneously.

Our findings have several implications. First, we have seen that the propensity for extreme categories was associated with higher variability in affect measures. In addition, we saw descriptively that those participants who showed high ERS in a questionnaire that was filled out at the end of the study also displayed a higher propensity to respond in the extreme categories in their affective responses throughout the experiment (see Figure 9). This result implies that a part of the variability in ILD may not be due to "true" variability in the construct that is measured. This is an important observation, as ERS has been shown to be consistent across content domains and over time (Van Vaerenbergh & Thomas, 2013; Weijters et al., 2010a; Wetzel et al., 2013), meaning our results may not be limited to the affective domain, but that ERS is likely to impact data that we obtain through our measures in various ILD domains in psychological and clinical research.

Interestingly, however, we did not find evidence that a participant's ERS influences the conclusions one reaches through the use of a typical autoregressive or multilevel model. In other words, neither inertia nor the slope of the monetary outcomes seems to be related to ERS. One explanation for this result may be that the potential range of statistical effects of ERS on these characteristics is small, as was also suggested for inertia in our simulated illustration presented in Figure 1.

Another interesting result pertains to the negative residual correlation that we observed between inertia and variability, which stands in contrast to previous studies that reported small to moderate positive correlations between these two measures (e.g., Bos et al., 2019; Wang et al., 2012). There might be several reasons for this difference in sign. One explanation might be that the skewed distribution of affect variables in previ-

ous studies exerted an impact on between-person correlation that was eliminated in our study with non-skewed distribution of affect variables. Another explanation might be the conceptual meaning of inertia, or the the moment-to-moment predictability, which might be different in real-world settings compared to an artificial experimental setup with short time interval between measurements (see also Bringmann et al., 2022, for a critical discussion on conceptual clarity).

Our results should be interpreted in light of several strengths and limitations. First, in contrast to previous studies (e.g., Baird et al., 2017) who used the same items to measure affect and ERS, we used a distinct background questionnaire to extract ERS indicators. This procedure allowed us to disentangle the measurement of affect dynamics from the measurement of ERS, which is corroborated by the low correlations between the ERS indicators and affect. Second, we induced affect levels in each experimental trial through varying the amount of wins and losses in a probabilistic reward task. As a consequence, affect measures in our study were approximately uniformly distributed without skewness and floor or ceiling effects, which – in contrast to previous studies – provided us with unconfounded measures of affect dynamics (see Baird et al., 2017; Bos et al., 2019; Deng et al., 2018; Koval et al., 2016). Lastly, this study was highly powered with $T = 140$ measurement occasions and $N = 1{,}398$ persons. This large sample allowed us to study affect variability, inertia, and reactivity to affect stimuli (see Wang et al., 2012).

While the experimental approach taken in this study yielded many advantages, it might also constrain the generalizability of results to real-world settings. It is therefore important that future research replicates our results in other suitable experimental and daily-life settings. Additionally, as the data were not primarily collected for the purpose

of this article, the sample characteristics of this study depend on the choices made by Vanhasbroeck et al. (2024). Unfortunately, they did not include several demographic variables that might have been informative with regard to generalizability to a broader population, of which we are therefore uncertain. Furthermore, while we did not find the the visual layout of the response scales had an impact on our results (see Appendix B and the supplementary materials on OSF), we cannot rule out that the association between ERS and affect dynamics might be different when using other scale formats. Thus, our study also points towards some interesting avenues for future research, such as investigating the influence of ERS in daily-life settings and varying the response and measurement formats. While we showed that variability was associated with ERS, there also is substantial residual variability that was not explained. Hence, affect variability may be an interesting subject of study in ILD and should be investigated further to learn more about interindividual processes and dynamics (e.g., Hedeker, Demirtas, & Mermelstein, 2009; Hedeker & Mermelstein, 2020; Wang et al., 2012).

More generally, the broad topics of measurement quality and measurement reliability have only recently gained more attention in ILD research (see e.g., Calamia, 2019; Cloos et al., 2023; Ringwald et al., 2022; Scott et al., 2020; Wright & Zimmermann, 2019). At the moment, standard procedures and clear guidelines how to evaluate data quality and control for potential confounding factors in ILD studies are missing. Our study has demonstrated that the magnitude of response biases may also vary between persons, corroborating the necessity to develop person-specific measures of data quality. Developing such standard procedures for self-report measures in ILD, but also passive measures are thus an essential next step to improve psychological and clinical assessments (Langener et al., 2024; Vize & Wright, 2024).

Lastly, modeling ERS using latent measurement models, such as item response theory or structural equation modeling, may oftentimes not be feasible for empirical researchers, for instance due to sample size restrictions. It would be desirable for everyday research practice to control for ERS using an approach which could be more easily integrated into the main analysis model. Future research may assess the effectiveness of such simplified approaches using manifest ERS trait levels that can be used in latent regressions as a control variable in contrast to more complex modeling approaches, and derive guidelines for statistical analyses of ILD data.

Assessing intraindividual variability has become more and more important in clinical studies (e.g., Jenkins et al., 2024; Mohr et al., 2015; Palmier-Claus et al., 2012; Russell et al., 2007). Our study is the first showing the association between response strategies and measures of affect variability in a controlled experimental setup. Recently, researchers have started to be interested in comparing measures of affect variability between regular and clinical samples or age groups, or use affect variability to detect transitions from regular to clinical states (e.g., Röcke et al., 2009; Schreuder et al., 2024; Trull et al., 2008). Our study provides the basis for developing more directed procedures to account and control for extreme responding and other response biases in measures of affect dynamics that will then allow researchers to conduct more robust analyses on clinically relevant affect dynamics. Our study is thus of foundational importance for all applied and clinical applications with measures of affect variability, and is a stepping stone to improved measures of affect dynamics in psychological and clinical assessments.

References

Adams, D. J., Bolt, D. M., Deng, S., Smith, S. S., & Baker, T. B. (2019). Using multidimensional item response theory to evaluate how response styles impact measurement. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 1–20. https://doi.org/10.1111/bmsp.12169

Asparouhov, T., & Muthen, B. (2010). *Bayesian analysis using Mplus: Technical implementation* (tech. rep.). https://statmodel.com/download/Bayes3.pdf.

Asparouhov, T., & Muthen, B. (2023). *Bayesian analysis using Mplus: Technical implementation* (tech. rep.). http://www.statmodel.com/download/Bayes2.pdf

Baird, B. M., Lucas, R. E., & Donnellan, M. B. (2017). The role of response styles in the assessment of intraindividual personality variability. *Journal of Research in Personality*, *69*, 170–179. https://doi.org/10.1016/j.jrp.2016.06.015

Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*, 665–678. https://doi.org/10.1037/a0028111

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*, 159–181. https://doi.org/10.1111/bmsp.12086

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*, 335–352. https://doi.org/10.1177/0146621608329891

Bolt, D. M., Lu, Y., & Kim, J.-S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*, *19*, 528–541. https://doi.org/10.1037/met0000016

Bos, E. H., de Jonge, P., & Cox, R. F. A. (2019). Affective variability in depression: Revisiting the inertia–instability paradox. *British Journal of Psychology*, *110*(4), 814–827. https://doi.org/10.1111/bjop.12372

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to the basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science*, *31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Brose, A., Schmiedek, F., Koval, P., & Kuppens, P. (2015). Emotional inertia contributes to depressive symptoms beyond perseverative thinking [Pub-

lisher: Taylor & Francis]. *Cognition and Emotion*, *29*(3), 527–538. https://doi.org/10.1080/02699931.2014.916252

Calamia, M. (2019). Practical considerations for evaluating reliability in ambulatory assessment studies. *Psychological Assessment*, *31*(3). https://doi.org/10.1037/pas0000599

Carver, C. S. (2015). Control processes, priority management, and affective dynamics. *Emotion Review*, *7*(4), 301–307. https://doi.org/10.1177/1754073915590616

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Cloos, L., Ceulemans, E., & Kuppens, P. (2023). Development, validation, and comparison of self-report measures for positive and negative affect in intensive longitudinal research. *Psychological Assessment*, *55*(3), 189–204. https://doi.org/10.1037/pas0001200

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology*, *60*, 151–174. https://doi.org/10.1037/h0040372

Czyz, E. K., Yap, J. R., King, C. A., & Nahum-Shani, I. (2021). Using intensive longitudinal data to identify early predictors of suicide-related outcomes in high-risk adolescents: Practical and conceptual considerations. *Assessment*, *28*(8), 1949–1959. https://doi.org/10.1177/1073191120939168

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*, 1–28. https://doi.org/10.18637/jss.v048.c01

De Haan-Rietdijk, S., Kuppens, P., & Hamaker, E. L. (2016). What's in a day? A guide to decomposing the variance in intensive longitudinal data. *Frontiers in Psychology*, *7*, 1–16. https://doi.org/10.3389/fpsyg.2016.00891

De Jong, M. G., Steenkamp, J.-B. E., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*, 104–115. https://doi.org/10.1509/jmkr.45.1.104

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

Deng, S., McCarthy, D. E., Piper, M. E., Baker, T. B., & Bolt, D. M. (2018). Extreme response style and the measurement of intra-individual variability in affect [Publisher: Taylor & Francis]. *Multivariate Behavioral Research*, *53*(2), 199–218. https://doi.org/10.1080/00273171.2017.1413636

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, *22*(2), 240–261. https://doi.org/10.1037/met0000065

Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology*, *76*(4), 662–676.

Ernst, A. F., Timmerman, M. E., Jeronimus, B. F., & Albers, C. J. (2021). Insights into individual differences in emotion dynamics with clustering. *Assessment*, *28*(4), 1186–1206. https://doi.org/10.1177/1073191119873714

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*, 328–347. https://doi.org/10.1037/met0000059

Franck, E., & De Raedt, R. (2007). Self-esteem reconsidered: Unstable self-esteem outperforms level of self-esteem as vulnerability marker for depression. *Behaviour Research and Therapy*, *45*(7), 1531–1541. https://doi.org/10.1016/j.brat.2007.01.003

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Guo, J., Marsh, H. W., Parker, P. D., Dicke, T., Lüdtke, O., & Diallo, T. M. O. (2019). A systematic evaluation and comparison between exploratory structural equation modeling and bayesian structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 529–556. https://doi.org/10.1080/10705511.2018.1554999

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, 1–18. https://doi.org/10.1080/10705511.2017.1402334

Hamaker, E. L., Grasman, R. P. P. P., & Kamphuis, J. H. (2016). Modeling BAS dysregulation in bipolar disorder: Illustrating the potential of time series analysis. *Assessment*, *23*(4), 436–446. https://doi.org/10.1177/1073191116632339

Hamaker, E. L., & Klugkist, I. (2011). Bayesian estimation of multilevel models. In *Handbook for Advanced Multilevel Analysis* (pp. 137–161). Routledge/Taylor & Francis Group.

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin*, *69*, 192–203. https://doi.org/10.1037/h0025606

He, J., & Van De Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Difference*, *55*, 794–800. https://doi.org/10.1016/j.paid.2013.06.017

He, J., & van de Vijver, F. J. R. (2016). Integration and domain specificity of response styles: Towards a better understanding of a general response style. *Psihologie Resurselor Umame*, *14*, 152–161.

Hedeker, D., Demirtas, H., & Mermelstein, R. J. (2009). A mixed ordinal location scale model for analysis of ecological momentary assessment (EMA) data. *Statistical Interface*, *2*(4), 391–401.

Hedeker, D., & Mermelstein, R. J. (2020). Modeling variation in intensive longitudinal data. In *Multilevel Modeling Methods with Introductory and Advanced Applications*. Information Age Publishing.

Hedeker, D., Mermelstein, R. J., Berbaum, M. L., & Campbell, R. T. (2009). Modeling mood variation associated with smoking: An application of a heterogeneous mixed-effects model for analysis of ecological momentary assessment (EMA) data. *Addiction*, *104*(2), 297–307. https://doi.org/10.1111/j.1360-0443.2008.02435.x

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location/scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, *64*(2), 627–634. https://doi.org/10.1111/j.1541-0420.2007.00924.x

Henninger, M. (2021). A novel Partial Credit extension using varying thresholds to account for response styles. *Journal of Educational Measurement*, *58*, 104–129. https://doi.org/10.1111/jedm.12268

Henninger, M., & Meiser, T. (2020a). Different approaches to modeling response styles in Divide-by-Total IRT models (Part I): A model integration. *Psychological Methods*, *25*, 560–576. https://doi.org/10.1037/met0000249

Henninger, M., & Meiser, T. (2020b). Different approaches to modeling response styles in Divide-by-Total IRT models (Part II): Applications and novel extensions. *Psychological Methods*, *25*, 577–595. https://doi.org/10.1037/met0000268

Henninger, M., & Meiser, T. (2022). Quality control: Response style modeling. In D. McCaffrey (Ed.), *International Encyclopedia of Education*. https://doi.org/10.1016/B978-0-12-818630-5.10041-7

Henninger, M., & Plieninger, H. (2021). Different styles, different times: How response times can inform our knowledge about the response process in rating scales. *Assessment*, *28*, 1–19. https://doi.org/10.1177/1073191119900003

Henninger, M., Plieninger, H., & Meiser, T. (2023). The effect of response formats on response style strength: An experimental comparison. *European Journal of Psychological Assessment*, 1–17. https://doi.org/10.1027/1015-5759/a000779

Hoffmann, S., Mai, R., & Cristescu, A. (2013). Do culture-dependent response styles distort substantial relationships? *International Business Review*, *22*(5), 814–827. https://doi.org/10.1016/j.ibusrev.2013.01.008

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*(4), 901–930. https://doi.org/10.1037/a0038822

Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment. *Psychological Methods*, *13*, 354–375. https://doi.org/10.1037/a0014173

Jenkins, B. N., Ong, L. Q., Ong, A. D., Lee, H. Y., & Boehm, J. K. (2024). Mean affect moderates the association between affect variability and mental

health. *Affective Science*, *5*(2), 99–114. https://doi.org/10.1007/s42761-024-00238-0

Jeon, M., & De Boeck, P. (2019). Evaluation on types of invariance in studying extreme response bias with an IRTree approach. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 517–537. https://doi.org/10.1111/bmsp.12182

Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, *49*, 161–177. https://doi.org/10.1080/00273171.2013.866536

Koval, P., Pe, M. L., Meers, K., & Kuppens, P. (2013). Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion*, *13*(6), 1132–1141. https://doi.org/10.1037/a0033579

Koval, P., Sütterlin, S., & Kuppens, P. (2016). Emotional inertia is associated with lower well-being when controlling for differences in emotional context. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01997

Kuppens, P., Van Mechelen, I., Nezlek, J. B., Dossche, D., & Timmermans, T. (2007). Individual differences in core affect variability and their relationship to personality and psychological adjustment. *Emotion*, *7*(2), 262–274. https://doi.org/10.1037/1528-3542.7.2.262

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology*, *17*, 22–26. https://doi.org/10.1016/j.copsyc.2017.06.004

Langener, A. M., Stulp, G., Jacobson, N. C., Costanzo, A., Jagesar, R. R., Kas, M. J., & Bringmann, L. F. (2024). It's all about timing: Exploring different temporal resolutions for analyzing digital-phenotyping data. *Advances in Methods and Practices in Psychological Science*, *7*(1), 1–22. https://doi.org/10.1177/25152459231202677

Larsen, R. J., & Diener, E. (1987). Affect intensity as an individual difference characteristic: A review. *Journal of Research in Personality*, *21*(1), 1–39. https://doi.org/10.1016/0092-6566(87)90023-7

Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling heterogeneous variance-covariance components in two-level models. *Journal of Educational and Behavioral Statistics*, *39*(5), 307–332. https://doi.org/10.3102/1076998614546494

Lin, X., Mermelstein, R. J., & Hedeker, D. (2018). A 3-level Bayesian mixed effects location scale model with an application to ecological momentary assessment data. *Statistics in Medicine*, *37*(13), 2108–2119. https://doi.org/10.1002/sim.7627

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1–16. https://doi.org/10.3389/fpsyg.2015.01171

Maher, J. P., Dzubur, E., Nordgren, R., Huh, J., Chou, C. P., Hedeker, D., & Dunton, G. F. (2019). Do fluctuations in positive affective and physical feeling states predict physical activity and sedentary time? *Psychology of*

Sport and Exercise, *41*, 153–161. https://doi.org/10.1016/j.psychsport.2018.01.011

McKone, K. M. P., & Silk, J. S. (2022). The emotion dynamics conundrum in developmental psychpathology: Similarities, distinctions, and adaptiveness of affective variability and socioaffective flexibility. *Clinical Child and Family Psychology Review*, *25*, 44–74. https://doi.org/10.1007/s10567-022-00382-8

McNeish, D. (2016). Using data-dependent prior to mitigate smalls ample bias in latente growth models: A discussion and illustration using Mplus. *Journal of Educational and Behavioral Statistics*, *41*(1), 27–56. https://doi.org/10.3102/1076998615621299

McNeish, D. (2021). Specifying location-scale models for heterogeneous variances as multilevel SEMs. *Organizational Research Methods*, *24*(3), 630–653. https://doi.org/10.1177/1094428120913083

Meiser, T., Plieninger, H., & Henninger, M. (2019). IRTree models with ordinal and multidimensional decision nodes for response styles and trait-based rating responses. *British Journal of Mathematical and Statistical Psychology*, *72*, 501–516. https://doi.org/10.1111/bmsp.12158

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

Mohr, C. D., Arpin, S., & McCabe, C. T. (2015). Daily affect variability and context-specific alcohol consumption. *Drug and Alcohol Review*, *34*(6), 581–587. https://doi.org/10.1111/dar.12253

Mourali, M., Böckenholt, U., & Laroche, M. (2007). Compromise and attraction effects under prevention and promotion motivations. *Journal of Consumer Research*, *34*(2), 234–247. https://doi.org/10.1086/519151

Muthen, B. (2010). Bayesian analysis in Mplus: A brief introduction. https://www.statmodel.com/download/IntroBayesVersion%203.pdf

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (2017). Mplus User's Guide.

Palmier-Claus, J. E., Taylor, P. J., Gooding, P., Dunn, G., & Lewis, S. (2012). Affective variability predicts suicidal ideation in individuals at ultra-high risk of developing psychosis: An experience sampling study. *British Journal of Clinical Psychology*, *51*(1), 72–83. https://doi.org/10.1111/j.2044-8260.2011.02013.x

Panksepp, J. (2012). What is an emotional feeling? Lessons about affective origins from cross-species neuroscience. *Motivation and Emotion*, *36*(1), 4–15. https://doi.org/10.1007/s11031-011-9232-y

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality*

and Social Psychological Attitudes (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Piasecki, T. M., Hedeker, D., Dierker, L. C., & Mermelstein, R. J. (2016). Progression of nicotine dependence, mood level, and mood variability in adolescent smokers. *Psychology of Addictive Behaviors*, *32*(8), 859–860. https://doi.org/10.1037/adb0000429_2

Plieninger, H. (2017). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*, 32–53. https://doi.org/10.1177/0013164416636655

Plieninger, H. (2020). Developing and applying IR-Tree models : Guidelines, caveats, and an extension to multiple groups. *Organizational Research Methods*, 1–17. https://doi.org/10.1177/1094428120911096

Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, *53*, 633–654. https://doi.org/10.1080/00273171.2018.1469966

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*, 875–899. https://doi.org/10.1177/0013164413514998

Price, G. D., Heinz, M. V., Song, S. H., Nemesure, M. D., & Jacobson, N. C. (2023). Using digital phenotyping to capture depression symptom variability: Detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. *Translational Psychiatry*, *13*(1), 381. https://doi.org/10.1038/s41398-023-02669-y

Rast, P., Hofer, S. M., & Sparks, C. (2012). Modeling individual differences in within-person variation of negative and positive affect in a mixed effects location scale model using BUGS/JAGS. *Multivariate Behavioral Research*, *47*(2), 177–200. https://doi.org/10.1080/00273171.2012.658328

Ringwald, W. R., Manuck, S. B., Marsland, A. L., & Wright, A. G. C. (2022). Psychometric evaluation of a Big Five personality state scale for intensive longitudinal studies. *Assessment*, *29*(6), 1301–1319. https://doi.org/10.1177/10731911211008254

Röcke, C., & Brose, A. (2013). Intraindividual variability and stability of affect and well-being. *The Journal of Gerontopsychology and Geriatric Psychiatry*, *26*(3), 185–199. https://doi.org/10.1024/1662-9647/a000094

Röcke, C., Li, S.-C., & Smith, J. (2009). Intraindividual variability in positive and negative affect over 45 days: Do older adults fluctuate less than young adults. *Psychology & Aging*, *24*, 863–78. https://doi.org/10.1037/a0016276

Russell, J. J., Moskowitz, D. S., Zuroff, D. C., Sookman, D., & Paris, J. (2007). Stability and variability of affective experience and interpersonal behavior in borderline personality disorder. *Journal of Abnormal Psychology*, *116*(3), 578–588. https://doi.org/10.1037/0021-843X.116.3.578

Schreuder, M. J., Schat, E., Smit, A. C., Snippe, E., & Ceulemans, E. (2024). Monitoring emotional intensity and variability to forecast depression re-

currence in real time in remitted adults. *Journal of Consulting and Clinical Psychology*. https://doi.org/10.1037/ccp0000871

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E. L. (2016). A comparison of inverse-wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, *51*(2-3), 185–206. https://doi.org/10.1080/00273171.2015.1065398

Scott, S. B., Sliwinski, M. J., Zawadzki, M., Stawski, R. S., Kim, J., Marcusson-Clavertz, D., Lanza, S. T., Conroy, D. E., Buxton, O., Almeida, D. M., & Smyth, J. M. (2020). A coordinated analysis of variance in affect in daily life. *Assessment*, *27*(8), 1683–1698. https://doi.org/10.1177/1073191118799460

Smit, A. C., Schat, E., & Ceulemans, E. (2023). The exponentially weighted moving average procedure for detecting changes in intensive longitudinal data in psychological research in real-time: A tutorial showcasing potential applications. *Assessment*, *30*(5), 1354–1368. https://doi.org/10.1177/10731911221086985

Trull, T. J., Lane, S. P., Koval, P., & Ebner-Priemer, U. W. (2015). Affective Dynamics in Psychopathology. *Emotion Review*. https://doi.org/10.1177/1754073915590617

Trull, T. J., Solhan, M. B., Tragesser, S. L., Jahng, S., Wood, P. K., Piasecki, T. M., & Watson, D. (2008). Affective instability: Measuring a core feature of borderline personality disorder with ecological momentary assessment. *Journal of Abnormal Psychology*, *117*(3), 647–661. https://doi.org/10.1037/a0012532

Ulitzsch, E., Henninger, M., & Meiser, T. (2024). Differences in response-scale usage are ubiquitous in cross-country comparisons and a potential driver of elusive relationships. *Scientific Reports*, *14*(1), 1–4. https://doi.org/10.1038/s41598-024-60465-0

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*, 195–217. https://doi.org/10.1093/ijpor/eds021

van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363–388. https://doi.org/10.1037/met0000162

Vanhasbroeck, N., Vanbelle, S., Moors, A., Vanpaemel, W., & Tuerlinckx, F. (2024). Chasing consistency: On the measurement error in self-reported affect in experiments. *Behavior Research Methods*, *56*(4), 3009–3022. https://doi.org/10.3758/s13428-023-02290-3

Vize, C. E., & Wright, A. G. C. (2024). Translating the transdiagnostic: Aligning assessment practices with research advances. *Assessment*, *31*(1), 199–215. https://doi.org/10.1177/10731911231194996

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*(3), 830–841. https://doi.org/10.1037/0033-295X.114.3.830

Wang, L., Hamaker, E., & Bergeman, C. (2012). Investigating inter-individual differences in short-term intra- individual variability. *Psychological Methods*, *17*, 19–44. https://doi.org/10.1037/a0029317

Weijters, B., Baumgartner, H., & Geuens, M. (2016). The calibrated sigma method: An efficient remedy for between-grop differences in response category use on Likert scales. *International Journal of Research in Marketing*, *33*(4), 944–960. https://doi.org/10.1016/j.ijresmar.2016.05.003

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, *34*, 105–121. https://doi.org/10.1177/0146621609338593

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods*, *15*, 96–110. https://doi.org/10.1037/a0018721

Wendt, L. P., Wright, A. G., Pilkonis, P. A., Woods, W. C., Denissen, J. J., Kühnel, A., & Zimmermann, J. (2020). Indicators of affect dynamics: Structure, reliability, and personality correlates. *European Journal of Personality*, *34*(6), 1060–1072. https://doi.org/10.1002/per.2277

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*, *33*, 352–364. https://doi.org/10.1027/1015-5759/a000291

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*, 178–189. https://doi.org/10.1016/j.jrp.2012.10.010

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*, 279–291. https://doi.org/10.1177/1073191115583714

Wichers, M., Wigman, J. T. W., & Myin-Germeys, I. (2015). Micro-level affect dynamics in psychopathology viewed from complex dynamical system theory. *Emotion Review*, *7*(4), 362–367. https://doi.org/10.1177/1754073915590623

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org/

Wright, A. G. C., & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological Assessment*, *31*(12), 1467–1480. https://doi.org/10.1037/pas0000685

A Questionnaire assessing how participants experienced the experiment.

Table 4: Questions used by Vanhasbroeck et al. (2024) to assess how participants experienced the experiment and the rating scale formats used to assess their affective states.

Order	Question
Q1	The experiment elicited positive feelings.
Q2	The experiment elicited negative feelings.
Q3	I felt indifferent about the experiment.
Q4	Winning or losing money did not have any effect on me.
Q5	I was motivated to finish the experiment.
Q6	The experiment was boring.
Q7	The experiment was frustrating.
Q8	I enjoyed the experiment.
Q9	The experiment was over sooner than I thought.
Q10	The emotion measure was easy to use.
Q11	I quickly understood how to use the emotion measure.
Q12	It was difficult to report on my feelings.
Q13	The emotion measure was confusing to use.
Q14	I was able to accurately describe my feelings with the emotion measure.
Q15	When my feelings changed, I could describe these changes accurately with the emotion measure.
Q16	My responses on the emotion measure conveyed how I really felt during the experiment.
Q17	My feelings could not be adequately captured by the emotion measure.

B Robustness checks of the data analysis models

Supplementary materials to these robustness checks are presented on OSF.

B.1 IRTree model

First, we assessed the descriptive person-level correlations of sumscores of the pseudo-items with person means of affect. We expected these correlation to be very low, because they stem from distinct questionnaires and measure distinct characteristics. Indeed, the person means of the extremity pseudo-items were essentially uncorrelated to the person means of affect measures (positive affect: $r = 0.04$; negative affect: $r = -0.04$; valence: $r = 0.07$). These low correlations indicate that our measure of extreme responding is not linearly associated with the measures of affect, and that we are able to distinguish extreme responding and affect as expected.

Second, we conducted a series of robustness checks for the IRTree model: We compared extremity estimates of a single extremity factor model (similar to Henninger & Plieninger, 2021) to the IRTree model. In addition, we compared an IRTree model with more than one agreement dimensions to the classical IRTree model to account for potential content multidimensionality or facets. All models showed comparable extremity trait estimates with correlations between the extremity process of $r = .99$, indicating that the estimation of the extremity trait factor was very stable across different model specification.

B.2 Impact of the visual layout of the response scales

We conducted robustness analyses for the effect of the visual layout of the response scales (continuity and rating scale format; see Figure 2) on the main results. For this purpose, we included main effects of the factors `continuity`

(continuous vs. discrete), and `format` (1D VAS, ESG, 2D VAS) on location and scale (i.e., variability) of positive and negative affect, as well as valence in the location scale model. The Mplus output files can be assessed on OSF. In summary, the robustness checks indicated that our results are largely unaffected.

B.3 Prior sensitivity analyses

The inverse-Wishart prior for the covariance matrix is beneficial, because it ensures that the diagonal follows an inverse gamma distribution and that the covariance matrix is positive definite. However, it can be difficult to specify uninformative prior distributions, in particular when the variances are small.

To investigate this case, we conducted prior sensitivity analyses for the inverse-Wishart prior. We assessed an improper inverse-Wishart prior $IW(1, -4)$ and $IW(0, -4)$, for variances and covariances, respectively, and a data-dependent prior using variance and covariance estimates of the IRTree model estimated using marginal maximum likelihood multiplied by the degrees of freedom ($p+1 = 4$; see Depaoli & van de Schoot, 2017; McNeish, 2016; Schuurman et al., 2016; van Erp et al., 2018, for more details and discussions).

For the location-scale model, instead of the improper inverse-Wishart prior, we also assessed proper inverse-Wishart priors, i.e., $IW(1, 5)$, $IW(0, 5)$ and $IW(1, 4)$, $IW(0, 4)$, for variances and covariances, respectively, varying the degrees of freedom ($p = 4$, and $p + 1 = 5$). We observed consistent parameter estimates across prior specifications, indicating that model estimates are robust against the prior specifications which would also be expected given the large sample size in our study (cf. Hamaker & Klugkist, 2011).

C Regression coefficients of the location-scale model

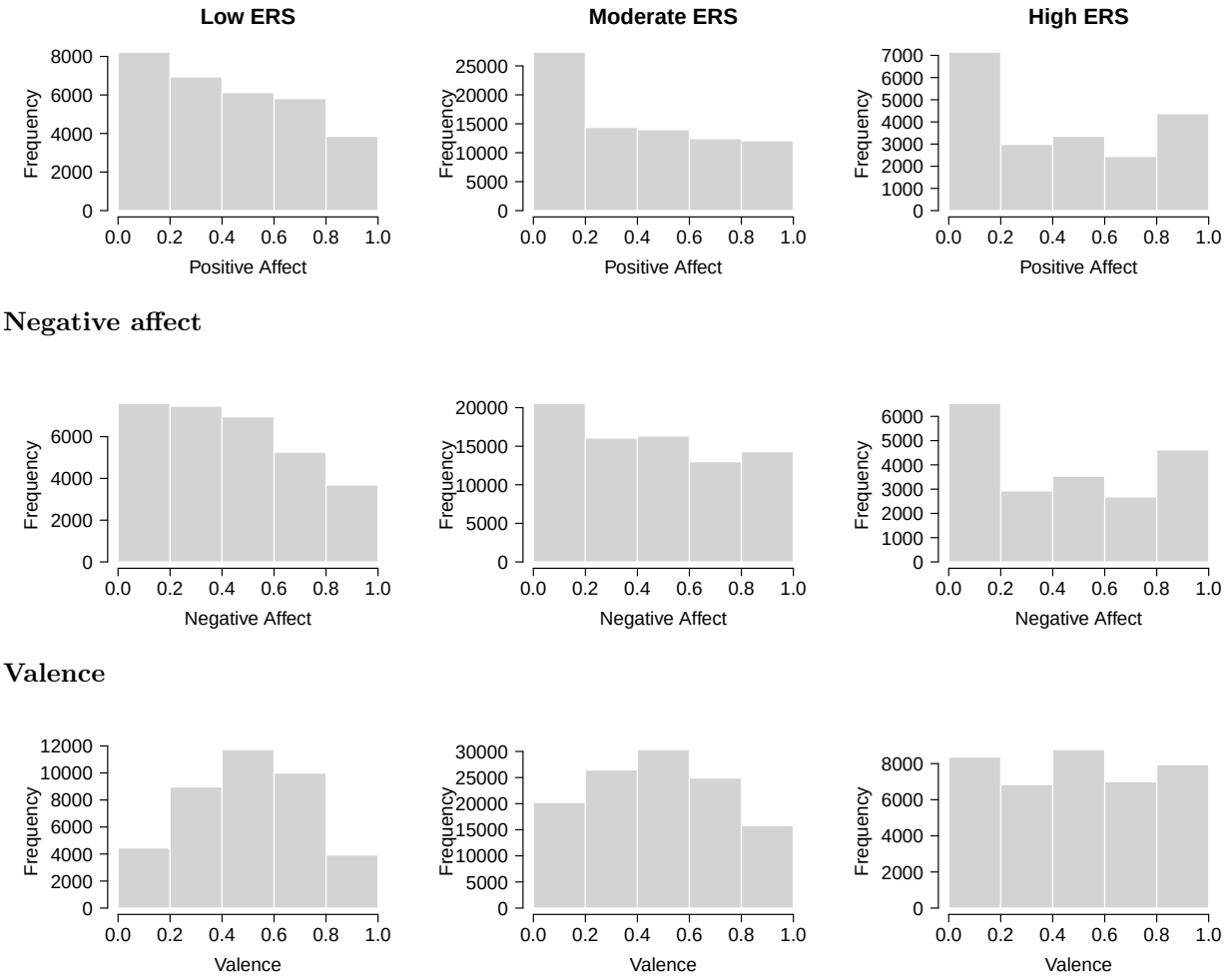
Table 5: Positive Affect

Effect	Notation	Posterior Median	95% Credible Interval
Location fixed effects			
Intercept	γ_{00}	0.40	[0.39, 0.41]
Inertia	γ_{10}	0.26	[0.24, 0.28]
Outcome	γ_{20}	0.58	[0.56, 0.60]
θ^{ERS}	γ_{03}	-0.01	[-0.03, 0.01]
Inertia $\times\theta^{\text{ERS}}$	γ_{13}	0.01	[-0.01, 0.05]
Outcome $\times\theta^{\text{ERS}}$	γ_{23}	0.01	[-0.02, 0.05]
Scale fixed effects			
ln(Intercept)	ω_0	-4.32	[-4.37, -4.27]
ln(θ^{ERS})	ω_1	0.09	[0.004, 0.18]
Random effects: Residual variances			
Random intercept	τ_{00}	0.02	[0.02, 0.02]
Random slope Inertia	τ_{11}	0.07	[0.06, 0.08]
Random slope Outcome	τ_{22}	0.10	[0.09, 0.11]
Person-level scale variance	τ_{33}	0.54	[0.49, 0.59]
Random effects: Correlations			
Intercept - Inertia	τ_{01}	-0.20	[-0.26, -0.14]
Intercept - Outcome	τ_{02}	0.31	[0.23, 0.39]
Intercept - Scale	τ_{03}	0.15	[0.08, 0.22]
Inertia - Outcome	τ_{12}	-0.70	[-0.73, -0.66]
Inertia - Scale	τ_{13}	-0.37	[-0.42, -0.32]
Outcome - Scale	τ_{23}	0.66	[0.59, 0.72]

Table 6: Negative Affect

Effect	Notation	Posterior Median	95% Credible Interval
Location fixed effects			
Intercept	γ_{00}	0.49	[0.48, 0.50]
Inertia	γ_{10}	0.30	[0.29, 0.32]
Outcome	γ_{20}	-0.49	[-0.52, -0.47]
θ^{ERS}	γ_{03}	0.02	[0.01, 0.04]
Inertia $\times\theta^{\text{ERS}}$	γ_{13}	0.02	[-0.01, 0.06]
Outcome $\times\theta^{\text{ERS}}$	γ_{23}	-0.04	[-0.07, -0.001]
Scale fixed effects			
ln(Intercept)	ω_0	-4.20	[-4.25, -4.15]
ln(θ^{ERS})	ω_1	0.10	[0.01, 0.19]
Random effects: Residual variances			
Random intercept	τ_{00}	0.02	[0.02, 0.03]
Random slope Inertia	τ_{11}	0.08	[0.07, 0.09]
Random slope Outcome	τ_{22}	0.10	[0.09, 0.10]
Person-level scale variance	τ_{33}	0.61	[0.56, 0.67]
Random effects: Correlations			
Intercept - Inertia	τ_{01}	0.23	[0.16, 0.31]
Intercept - Outcome	τ_{02}	-0.11	[-0.18, -0.04]
Intercept - Scale	τ_{03}	0.17	[0.10, 0.24]
Inertia - Outcome	τ_{12}	0.61	[0.56, 0.65]
Inertia - Scale	τ_{13}	-0.38	[-0.44, -0.32]
Outcome - Scale	τ_{23}	-0.45	[-0.50, -0.39]

Figure 9: Affect response distribution for low, medium, and high extreme responding (ERS) levels.



Note: The affect response distribution for positive and negative affect visually differs from the distribution for valence. Because valence is a composite score of positive and negative affect (see Equation 1), the distribution of valence measures is approximately normal compared to the distribution of positive and negative affect, with more responses in the center of the scale (around 0.5). As a consequence, the distribution under high ERS (rightmost column) also shows more responses in the center of the scale for valence compared to positive and negative affect.

Table 7: Valence

Effect	Notation	Posterior Median	95% Credible Interval
Location fixed effects			
Intercept	γ_{00}	0.46	[0.45, 0.46]
Inertia	γ_{10}	0.28	[0.26, 0.29]
Outcome	γ_{20}	0.57	[0.55, 0.58]
θ^{ERS}	γ_{03}	-0.01	[-0.02, -0.001]
Inertia $\times\theta^{\text{ERS}}$	γ_{13}	0.01	[-0.01, 0.04]
Outcome $\times\theta^{\text{ERS}}$	γ_{23}	0.04	[0.01, 0.07]
Scale fixed effects			
ln(Intercept)	ω_0	-4.80	[-4.84, -4.76]
ln(θ^{ERS})	ω_1	0.13	[0.05, 0.21]
Random effects: Residual variances			
Random intercept	τ_{00}	0.01	[0.01, 0.01]
Random slope Inertia	τ_{11}	0.08	[0.07, 0.09]
Random slope Outcome	τ_{22}	0.09	[0.09, 0.10]
Person-level scale variance	τ_{33}	0.55	[0.51, 0.60]
Random effects: Correlations			
Intercept - Inertia	τ_{01}	-0.28	[-0.34, -0.22]
Intercept - Outcome	τ_{02}	0.08	[0.01, 0.14]
Intercept - Scale	τ_{03}	-0.12	[-0.18, -0.07]
Inertia - Outcome	τ_{12}	-0.68	[-0.71, -0.65]
Inertia - Scale	τ_{13}	-0.35	[-0.40, -0.30]
Outcome - Scale	τ_{23}	0.50	[0.45, 0.55]