Empirical Paper

EUROPEAN JOURNAL OF Personality

European Journal of Personality
2025, Vol. 39(1) 55–69

DOI: 10.1177/08902070241235969
journals.sagepub.com/home/ejop
Sage

If you were happy and you know it, clap your hands! Testing the peak-end rule for retrospective judgments of well-being in everyday life

Julian Scharbert[1], Katharina Utesch[1], Thomas Reiter[2], Julian ter Horst[3], Maarten van Zalk[3], Mitja D Back[1,4] and Richard Rau[1,5]

Abstract

The experience sampling method (ESM) and comparable assessment approaches are increasingly becoming popular tools for well-being research. In part, they are so popular because they represent more direct approaches for assessing individuals' experienced well-being during a specified period, whereas one-time, retrospective evaluations of that episode are believed to introduce systematic biases. Along these lines, the *peak-end rule* states that the most extreme and recent sensations of an episode disproportionally influence retrospective well-being judgments. However, it has yet to be determined whether such systematic effects found in experimental laboratory studies generalize to retrospective judgments of well-being in everyday life as captured in ESM studies. Across four ESM samples (overall N = 1,889, total measurements = 131,575), we found that retrospective well-being judgments were disproportionately influenced by the peak and end experiences from the assessment period. However, these effects depended on the item framing of the retrospective judgment (global vs. more specific framings) and broad versus narrow conceptualizations of peaks and ends (states, days, and weeks), pointing toward potential ways to mitigate peak/end effects. Our findings emphasize the importance of differentiating between momentary and retrospective well-being assessments and selecting an appropriate measurement approach on the basis of these conceptual considerations.

Plain language summary

In psychology, there is a prevailing notion that each of us has two distinct selves: our experiencing self and our remembering self. While the former reflects our actual perceptions, emotions, and behaviors in the moment, the latter reflects our memories of these perceptions, emotions, and behaviors in hindsight. Crucially, these two aspects of our self do not always align perfectly, but there are systematic discrepancies between the two. For example, the peak-end rule asserts that when individuals recall their past experiences, they primarily rely on the peak (i.e., the best or worst moment) and the end (i.e., the last moment) of those experiences. The peak-end rule has been demonstrated primarily in the laboratory, but evidence has been mixed as to whether it also holds in everyday life, that is, whether evaluations of our past well-being are disproportionally influenced by peak and end experiences. In this study, we show that our peak and end experiences in a given episode can have a disproportionate influence on how happy we think we were. However, we also show that the extent of this influence depends on how we ask people to recall their well-being (e.g., whether we ask them to evaluate their well-being in general versus whether we ask them to refer to specific moments) and on how broadly we frame peaks and ends (e.g., peak/end weeks have stronger influences than specific peak/end moments). Overall, our study underscores that our memories of how happy we were do not always resemble how happy we actually were and that the degree of such deviations depends on the framing of the remembering self. Psychologists should keep that in mind when assessing well-being.

Keywords

well-being, peak-end rule, experience sampling, retrospective judgments, cognitive biases

Received 15 May 2023; Revised 28 November 2023; accepted 9 February 2024

In recent decades, the *experience sampling method* (ESM; Csikszentmihalyi & Larson, 1987) and comparable methods, such as the *daily diary method* and *day reconstruction method* (Kahneman et al., 2004), have become more and more popular tools for researchers who are interested in well-being (Conner & Barrett, 2012; Wilhelm et al., 2012; Wrzus & Neubauer, 2023). Besides other advantages, these methods are considered to represent a less

[1]Institute of Psychology, University of Münster, Münster, Germany
[2]Department of Psychology, University of Munich, Munich, Germany
[3]Institute of Psychology, Osnabrück University, Osnabruck, Germany
[4]Joint Institute for Individualisation in a Changing Environment (JICE), Bielefeld University, University of Münster, Münster, Germany
[5]Health and Medical University Potsdam, Potsdam, Germany

Corresponding author:
Julian Scharbert, Institute of Psychology, University of Münster, Fliednerstr. 21, Münster 48149, Germany.
Email: julian.scharbert@uni-muenster.de

biased measurement approach for assessing well-being because they enable researchers to access aspects of individuals' moment-to-moment reality that cannot be fully recaptured in hindsight. Accordingly, asking individuals just once (one-time, trait-level report) about their well-being over a certain period of time should not be fully representative of their momentary experiences (i.e., state-level reports; Schwarz, 2012). Indeed, whereas mean scores of multiple momentary assessments are related to trait scores to some degree, correlations are far from perfect. For instance, correlations between momentary and retrospective well-being judgments typically range from .40 to .80 (e.g., Ben-Zeev et al., 2009; Cohen et al., 1995; Kritzler et al., 2020; Scollon et al., 2009; Wilt et al., 2012). Such divergences between momentary and retrospective reports can have many *unsystematic* reasons, such as the imperfect reliability of assessment methods or the "random" forgetting of experiences that introduces noise into the retrospective retrieval of these experiences. However, there can also be *systematic* reasons for these divergences that can make inferences from one-time, retrospective assessments to momentary experiences (and vice versa) problematic.

According to the *peak-end rule* (Fredrickson, 2000; Kahneman et al., 1993), two of the most powerful sources of systematic biases in retrospective judgments are the most extreme and the most recent experience within a given episode. That is, when individuals form retrospective judgments, they primarily rely on the peak (i.e., the best or worst moment) and the end (i.e., the last moment) of an episodic experience. Initial empirical support for the peak-end rule stems from experimental studies in laboratory settings (e.g., Fredrickson & Kahneman, 1993; Kahneman et al., 1993; Redelmeier & Kahneman, 1996). Importantly, recent studies have suggested that the peak-end rule also holds for retrospective well-being judgments in data obtained outside the laboratory (e.g., Geng et al., 2013; Neubauer et al., 2020; Parkinson et al., 1995). However, evidence has been mixed, as additional studies have found either only peak or only end effects on retrospective well-being judgments (e.g., Ganzach & Yaor, 2019) or no significant effects of either peaks or ends after controlling for the mean across all momentary assessments (e.g., Ben-Zeev et al., 2009; Röcke et al., 2011).

The goal of this study is to clarify whether and, if so, how individuals' retrospective well-being judgments systematically deviate from their momentary well-being assessments in everyday life. Specifically, we aim to investigate whether the peak-end rule applies to one-time global and retrospective judgments of well-being in experience-sampled data (see https://osf.io/32b8a/ for our preregistration). For this purpose, we relied on four large ESM data sets (total N = 1,889, total measurements = 131,575), compared different framings of retrospective well-being judgments, and explored time intervals of different lengths to identify the peaks and ends in individuals' state data.

Momentary versus retrospective well-being judgments

One-time, retrospective judgments comprise global judgments (e.g., "How do you feel in general?) and judgments referring to a specific period (e.g., "How have you felt during the past month?"). Whereas global judgments require people to remember and aggregate their experiences over the widest possible time frame (i.e., one's life up to this point), explicit retrospective judgments narrow down the time frame (e.g., to the past day, week, month, or year). Both types of self-reports get applied frequently in psychological research and practice. For example, one of the classic versions of the *Positive and Negative Affect Schedule* (PANAS; Watson et al., 1988) asks participants about the extent to which they feel certain affective states "in general." By contrast, other scales, particularly those applied in the clinical context, ask participants to indicate their experiences in the past week (e.g., SCL-90-R; Hessel et al., 2001), the past 2 weeks (e.g., PHQ-9; Kroenke et al., 2001),[1] or the past month (e.g., OCI-R; Foa et al., 2002).

Importantly, because it might be cognitively too challenging to recall all details from an affective episode, it is believed that individuals tend to rely on especially salient aspects of the episode to evaluate it in hindsight (Conner & Barrett, 2012; Robinson & Clore, 2002). Two aspects of an episode seem to be particularly influential on retrospective judgments: the most intense (i.e., the "peak") and the most recent (i.e., the "end") moments from an episode. This notion emerged from studies that have investigated retrospective judgments in laboratory contexts. For example, Fredrickson and Kahneman (1993) found that retrospective evaluations of emotional film clips could be best predicted by a combination of the most extreme affect rating and the affect rating of the final moments. Analogously, Kahneman et al. (1993) and Redelmeier and Kahneman (1996) found that participants were more willing to endure a painful experience again (in this case: a colonoscopy or submerging one's hand in very cold water) if the episode had a slightly less painful ending, even if that meant extending the overall duration of the painful experience. These first empirical studies led to the formulation of the *peak-end rule* (e.g., Fredrickson, 2000), which asserts that retrospective judgments are disproportionally influenced by the peak and end moments of the experience. The peak-end rule has since been investigated in various contexts, such as vacations (Geng et al., 2013), relationship satisfaction (Zygar-Hoffmann & Schönbrodt, 2020), or business funding (Jiang et al., 2019), and a recent meta-analysis by Alaybek et al. (2022) indicated that its effect on retrospective evaluations is large and robust across contexts. Thus, the peak-end rule is a strong candidate for better understanding discrepancies between retrospective and momentary judgments.

Whereas insights generated by these studies are important for predicting various outcomes (e.g., life quality ratings, Diener et al., 2001; willingness to undergo a medical operation, Redelmeier et al., 2003), research is needed to clarify whether and, if so, to what degree they can be applied to retrospective well-being judgments on data obtained outside the laboratory. Initial studies have addressed this question but have yielded rather inconclusive findings: Whereas some of these studies found significant effects of the peak, the end, or both in momentary well-being on retrospective well-being judgments (Ganzach & Yaor, 2019; Geng et al., 2013; Neubauer et al., 2020; Parkinson et al., 1995), others did not (Ben-Zeev et al.,

2009; Röcke et al., 2011). These inconsistent results might be partly explained by the heterogeneity of the applied time frames for associating momentary with retrospective well-being data, with studies linking momentary to daily (Ganzach & Yaor, 2019; Neubauer et al., 2020; Parkinson et al., 1995), weekly (Ben-Zeev et al., 2009; Geng et al., 2013) or even yearly retrospective well-being judgments (Röcke et al., 2011). Additionally, many previous studies have suffered from methodological limitations: First, most studies included small samples with fewer than 80 participants (Ben-Zeev et al., 2009; Geng et al., 2013; Parkinson et al., 1995; Röcke et al., 2011). Thus, they lacked the statistical power to find smaller effects of peaks and ends beyond average affect ratings. Second, when large and significant associations between retrospective well-being judgments and momentary peaks and ends were found, they had to be interpreted with caution if the authors had not controlled for the effect of the average momentary ratings (e.g., Geng et al., 2013). As peaks and ends have typically been found to be strongly correlated with the average level of momentary assessments (e.g., correlations ranged from .60 to .69 for ends and from .79 to .84 for peaks in Neubauer et al., 2020; Parkinson et al., 1995), not accounting for these associations makes it difficult to determine whether or not there is an overreliance on peaks or ends. Finally, momentary judgments were sometimes operationalized as daily reports (Ganzach & Yaor, 2019; Geng et al., 2013), which are already retrospective in nature. To rule out the possibility that some aggregation or interpretation processes might have already occurred and influenced the daily retrospective judgments, a more direct approach is to assess well-being in the moment.

The role of item framings

Item framings, such as the wording of the item or the context in which it is presented, play an important role in eliciting evaluation processes and decisions in participants (Kam, 2018; Schwarz, 1999; Tversky & Kahneman, 1981). This effect is relevant to the research question at hand, as there are different ways to prompt retrospective well-being judgments that, for instance, might entice participants to spend more time forming their responses or to access different kinds of information to arrive at their conclusions. Therefore, we obtained and compared three kinds of retrospective judgments from participants (Prompts 1–3). These three item framings vary in the applied time frame and specificity, and one can derive different hypotheses about how this affects their relationships to momentary well-being judgments on the basis of two theoretical frameworks (see Table 1).

On the one hand, one might form an argument that is in line with *dual-process theories* of cognition (e.g., Evans, 2003; Evans & Stanovich, 2013), which have been linked explicitly with the literature on the peak-end rule by Kahneman and colleagues (e.g., Kahneman, 2011; Kahneman & Frederick, 2012). These theories propose that there are two distinct kinds of human reasoning, often labeled *System 1* and *System 2* (e.g., Stanovich & West, 2000). System 1 is characterized by unconscious and intuitive reasoning, and people are more likely to make use of heuristics and automatic responses to arrive at decisions when using System 1. System 2 is characterized by more deliberate and analytical reasoning and is less prone to biases. More familiar and unconscious tasks elicit faster System 1 operations, whereas tasks that are novel or require deliberate thought elicit slower System 2 operations. On the basis of this reasoning, one would assume that enticing participants to think more deeply about their responses (i.e., engaging in System 2 rather than System 1 operations) should reduce their proclivity to rely on heuristics, such as using peaks and ends as a proxy for their experience (see Table 1).

On the other hand, one might form an argument that is based on the *accessibility model of emotional self-report* (Robinson & Clore, 2002), thus resulting in different predictions. According to this model, emotional reports that refer to a more specific episode or to a shorter time frame are more likely to engage episodic memory processes (i.e., retrieving one's memories by asking oneself: "How did I feel in this episode?"). By contrast, less emotional reports relating to longer or undefined time frames are thought to rely less on episodic information but instead rely on semantic information (i.e., retrieving one's self-concept by asking oneself: "How do I usually feel?"). Thus, one might expect more specific prompts to entice participants to use episodic memory to recreate the episodic details of their experiences, making it more likely that episodic memory biases will be introduced, such as peak and end effects (see Table 1). As both lines of reasoning seemed plausible to us, we refrained from forming concrete hypotheses about effects of item framings but instead addressed this issue in an exploratory manner.

The present research

Here, we provide a comprehensive investigation of the potential discrepancies of retrospective and global judgments from the mean of experience-sampled state data in well-being pertaining to the peak-end rule (Fredrickson, 2000). We address methodological shortcomings and inconsistencies in prior research by (a) testing our preregistered hypotheses (see https://osf.io/32b8a/) and exploratory research questions in four ESM samples (overall N = 1,889), (b) investigating peak and end effects beyond mean-level effects, (c) linking retrospective to momentary rather than daily well-being judgments, (d) exploring time intervals of different lengths to identify the peaks and ends in individuals' state data (i.e., peak/end moments, days, and weeks), and (e) comparing different item framings for retrospective judgments.

Method

We analyzed data collected as part of the multistudy project "Coping with Corona" (Scharbert et al., 2023). Specifically, our analysis involved two samples from an international ESM study (Sample 1, German subsample; Sample 2, Polish subsample) and two related German ESM studies (Samples 3 and 4). Because the samples were comparable in many regards, we describe the participants, procedures, and measures jointly for all samples. We focus on the parts of

Table 1. Predicted strengths of peak/end effects by item framing.

Item framing	Predicted by dual-process theories of cognition	Predicted by accessibility model of emotional self-report
Prompt 1: *"In general, I feel happy."*	Strongest effects of peak and end (response is expected to be based primarily on System 1 processing)	Weakest effects of peak and end (response is expected to be based primarily on semantic information)
Prompt 2: *"Over the past 4 weeks, I felt happy."*	In-between strength of effects	In-between strength of effects
Prompt 3: *"Over the past 4 weeks, you have regularly indicated how you felt. What would you say was your average response to the statement 'I feel happy'?"*	Weakest effects of peak and end (response is expected to be based primarily on System 2 processing)	Strongest effects of peak and end (response is expected to be based primarily on episodic information)

the studies with direct relevance to the present investigation. Codebooks with a full description of all procedures and measures applied in this project as well as the data analytic codes and the data needed to reproduce the presented results can be retrieved from https://osf.io/32b8a/. All procedures from these studies were approved by the review board of the University of Münster (2020-54-MB).

Procedures

Both the international ESM study (Samples 1 and 2) and the two German ESM studies (Samples 3 and 4) consisted of three phases: a pre-survey, a 4-week ESM period, and a post-survey. Upon registration, participants completed the pre-survey, in which we assessed demographic and several other trait-level variables not considered here. Throughout the following 4-week ESM period, participants were contacted at four random time points per day between 9 am and 6 pm (Samples 1 and 2) or at five random time points per day between 7 am and 6:30 pm[2] (Samples 3 and 4) and asked to complete a brief survey in which they indicated their momentary well-being. After 4 weeks, participants completed the post-survey, in which we assessed retrospective and global well-being judgments.

The international ESM study was programmed completely in the *formr* software (Arslan et al., 2020) version v0.18.3. Invitational links to the brief ESM surveys were sent out via email. For the two German ESM studies, the pre- and post-surveys were programmed in the *Unipark* software (Tivian, 2022) version 22. Depending on a participant's phone's operating system, the brief surveys were delivered either via a smartphone research app upon phone activation (Android) or via email and the online survey software *SoSci Survey* (iOS; Leiner, 2019) version 3.4.13. We merged the data collected via SoSci Survey and the research app after the study had been completed.

Because the three German samples were collected by the same team of researchers, a minority of participants were included in multiple data sets. To account for this fact, we conducted robustness analyses in which each participant was included only once (see Table A3 in the supplement).

Participants and data exclusion

The initial samples consisted of 1,092 (Sample 1), 376 (Sample 2), 227 (Sample 3), and 340 (Sample 4) participants who provided state data and at least one retrospective well-being judgment. Before including participants in our analyses, we applied the following preregistered data exclusion criteria: First, we excluded participants who indicated at the end of the study that they had responded carelessly to the items (five participants across samples) or who had completed the pre- or post-survey too quickly, that is, who took <2 s per item on average (twenty participants). Second, we excluded state data from all individual measurement occasions in which a participant had no variance across all the items on the survey page (all six state affect items), for example, if a participant always chose the items on the right-hand side of the scale (2,192 measurements). Third, to remove participants with only very limited well-being reports, we excluded participants who provided valid state data in fewer than 10 measurement occasions (89 participants). Finally, due to technical errors, the ESM period lasted for more than 28 days for some participants who occasionally skipped a day. Here, we excluded the most extreme cases in which participants had extended the ESM period for more than a week (32 participants). These data exclusions resulted in an overall sample of 1,889 participants (131,575 state measurements).

Samples 1 and 2 were part of the same international data collection project and represent the German and Polish subsamples, respectively. We selected these two subsamples because the German subsample represents the largest one, and in the Polish subsample, we assessed the two additional Prompts 2 and 3. Sample 1 consisted of 1,057 participants (892 women) between the ages of 18 and 91 ($M = 31.5$, $SD = 14.8$). Participants completed an average of 66.3 state assessments ($Mdn = 70$, range: 10–111), and the total number of state assessments was 70,104. Sample 2 consisted of 278 participants (176 women) between the ages of 19 and 82 ($M = 40.9$, $SD = 14.5$). Participants completed an average of 48.8 state assessments ($Mdn = 44$, range: 10–110), and the total number of state assessments was 13,561. Participants were recruited via social media, local and digital blackboards, mailing lists, university classes, recruitment panels, or local press releases. As incentives, participants received personalized feedback throughout, and after the data were collected, they could take part in a raffle for 10,000€ (prizes ranged from 20€ to 2,500€), and students in Germany could receive course credit. Furthermore, we donated 1€ per participant to one of three charity organizations that participants could select.

Samples 3 and 4 represent two sub-studies from the same data collection project conducted at three universities in Germany. Sample 3 consisted of 222 participants (187 women) between the ages of 18 and 47 (M = 21.4, SD = 3.9). Participants completed an average of 93.3 state assessments (Mdn = 97, range: 20–138), and the total number of state assessments was 20,706. Sample 4 consisted of 332 participants (282 women) between the ages of 17 and 48 (M = 21.4, SD = 4). Participants completed an average of 81.9 state assessments (Mdn = 86, range: 11–137), and the total number of state assessments was 27,204. In both samples, participants were recruited via announcements in university classes, local and digital blackboards, and semester groups on social media. As incentives, they were offered course credit and 30€ for complete participation. Android users received an additional 20€ for participating via the research app. Furthermore, if at least 80% of the students from a study cohort participated in the study, they were granted an additional 2,000€ for a communal event.

Measures

Momentary well-being judgments. Momentary well-being was assessed with the items "I felt happy" (Samples 1 and 2, referring to the last social interaction or individual activity in the past hour) or "I feel happy" (Samples 3 and 4, referring to the current moment), answered on a 6-point scale ranging from 1 (*do not agree at all*) to 6 (*agree completely*).

Retrospective well-being judgments. One-time well-being judgments were assessed in the post-survey after the 4-week ESM period with three types of retrospective judgments: In all the samples, participants rated a classic trait-level global statement about their well-being (Prompt 1: "In general, I feel happy") on a 5-point scale ranging from 1 (*very slightly or not at all*) to 5 (*extremely*). Participants in Samples 2, 3, and 4 additionally provided either a retrospective well-being judgment that referred specifically to the timespan during which the state assessments were obtained (Prompt 2: "Over the past 4 weeks, I felt happy") or a retrospective well-being judgment that referred explicitly to the average of their state assessments (Prompt 3: "Over the past 4 weeks, you have regularly indicated how you felt. What would you say was your average response to the statement 'I feel happy'?"). These judgments were made on a 6-point scale ranging from 1 (*do not agree at all*) to 6 (*agree completely*). The participants from Samples 2, 3, and 4 were randomly assigned to one of the two prompts (Prompt 2 or 3).

Other momentary and retrospective judgments. In supplementary analyses, we investigated whether peak and end effects are specific to retrospective well-being judgments or to what extent they can be generalized to other constructs. Thus, in addition to momentary and retrospective well-being judgments, we assessed participants' momentary and retrospective judgments of their anxiety (Prompt 1 in all samples, Prompts 2 and 3 only in Samples 2, 3, and 4), stress (no Prompt 1, Prompts 2 and 3 only in Samples 3 and 4), and energy level (no Prompt 1, Prompts 2 and 3 only in Sample 2) in a manner similar to the approach described above.[3]

Data-analytical approach

All analyses were conducted in *RStudio* (R Core Team, 2022) version 4.2.1. All data analytic codes and the data needed to reproduce the presented results can be retrieved from https://osf.io/32b8a/.

Calculation of means, peaks, and ends. For each participant, we calculated the mean of their state measurements. The end-point of the state data was the value of each participant's last assessment. The peak-point of the state data was the most positive value a participant had.[4] In addition, we explored different time intervals to determine the peaks and ends in the individuals' state data. Specifically, besides operationalizing the peak by the highest single state assessment score, we used the highest daily score (i.e., the mean across all state assessments on that day) and the highest weekly score (i.e., the mean across all state assessments in that week) as indicators of the peak of the state data. Analogously, we used the last day and the last week as indicators of the end of the state data.

Test of the peak-end rule. For each of the three types of retrospective judgments (Prompts 1–3), we initially predicted the retrospective judgments from the mean of the state assessments. These basic models were then extended by adding the peak and the end of the individual state assessments to the model as predictors. We then tested whether there was a significant difference in log likelihoods between the models that did and did not include both the peak and end of the state data as predictors. A significant difference between the models indicated that specific characteristics of the state data (i.e., the peak and end) disproportionally influenced retrospective judgments of momentary well-being, meaning that retrospective judgments did not provide an accurate representation of momentary well-being judgments. The hierarchical regression analyses using daily and weekly peaks and ends were conducted in the same manner.

Supplementary analyses. We conducted the following supplementary specificity and robustness analyses: First, we investigated whether the results of the main analyses generalized to retrospective judgments of anxiety, stress, and energy level and conducted the analyses described above with these variables instead of well-being. Second, we investigated whether the effects of endpoints could be accounted for by the effects of linear slopes by including the person-specific temporal slope of the well-being trajectories as a predictor of the retrospective well-being judgments in the models along with the peaks and ends of the state data. We then compared the predictive ability of these models with the person-specific slopes and the peaks and ends versus the models with only the peaks and ends of the state data. The temporal slope for each participant was obtained by calculating a multilevel model in which time as a Level-1 variable predicted momentary well-being across all state assessments of a participant and retrieving the random slope of that prediction for each participant. Finally, we investigated whether the results of the main analyses changed when we excluded participants who had participated in more than one sample.

Mega-analysis. In addition to investigating our research question in each sample separately, we applied a "mega-analytical" approach (Eisenhauer, 2021). That is, we included data from all the samples in one pooled sample and computed all the analyses described above in this pooled sample as well. To account for sample-specific effects, we additionally specified extended models in which we included the sample information as dummy-coded variables as predictors in the regression models as well as interaction terms between these sample dummies with the other predictors in the model (e.g., the mean). In these models, Sample 4 made up the reference category, meaning that the main effects of the means, peaks, and ends correspond to the effects in this sample rather than the overall main effects across samples. Therefore, we decided to focus on the more parsimonious models with the overall main effects here and report the results of the models that included the sample dummies and interaction terms in the supplement (Table A1).

Transparency and openness

Open science practices. We embrace the values of openness and transparency in science (Schönbrodt et al., 2015). We report all data exclusions and all measures in the study (Simmons et al., 2012) or refer to project documentation on the OSF. All data, analysis code, and research materials are available at https://osf.io/32b8a/. All hypotheses and the analysis plan were preregistered and can be found at https://osf.io/32b8a/. We did not conduct a priori power analyses. However, the sample sizes in all four samples were above average psychological ESM studies (around 136.6 in a meta-analysis by Wrzus and Neubauer (2023)). Thus, the statistical power we had to detect small effects was higher than usual, especially in the mega-analytical approach with N = 1,889 participants.

Prior use of data. The data from Samples 3 and 4 had not been used before in any other manuscripts published or in preparation. The data from Samples 1 and 2 were used in one manuscript that has been submitted for publication. That study did not link any momentary reports with retrospective reports and, thus, is unrelated to the present research and not described in detail here.

Results

Table 2 shows descriptive statistics and intercorrelations for the retrospective well-being judgments and the momentary well-being characteristics (means, peaks, and ends) in the pooled sample. The respective tables for each sample are presented separately in the supplement (Tables A9–A12). In general, the peaks and ends showed high correlations with the mean of the state well-being data. These associations increased for broader conceptualizations of peaks and ends (from $r \approx .60$ for peak/end states to $r \approx .90$ for peak/end weeks), meaning that the most positive and the last moment/day/week were increasingly indicative of the participants' overall well-being in the assessment period. In addition, a thorough inspection of the data revealed noteworthy differences across retrospective and momentary well-being judgments: The retrospective well-being judgments were significantly more positive than the mean of the momentary well-being judgments, d = 0.33, 95% CI [0.25, 0.42], $t(410) = 7.47$, $p < .001$ for Prompt 2 and $d = 0.35$, 95% CI [0.28, 0.41], $t(373) = 9.70$, $p < .001$ for Prompt 3.[5] Supplementary analyses revealed that these differences also emerged for retrospective and momentary anxiety judgments, $d = 1.20$, 95% CI [1.09, 1.31], $t(410) = 21.11$, $p < .001$ for Prompt 2 and $d = 0.67$, 95% CI [0.58, 0.76], $t(373) = 14.11$, $p < .001$ for Prompt 3; and for retrospective and momentary stress judgments, $d = 1.13$, 95% CI [0.99, 1.27], $t(204) = 16.04$, $p < .001$ for Prompt 2 and $d = 0.46$, 95% CI [0.34, 0.58], $t(215) = 7.78$, $p < .001$ for Prompt 3; but not for retrospective and momentary energy-level judgments, $d = -0.03$, 95% CI [−0.17, 0.10], $t(129) = -0.46$, $p = .649$ for Prompt 2 and $d = 0.12$, 95% CI [−0.02, 0.27], $t(100) = 1.68$, $p = .096$ for Prompt 3. These findings indicate that, overall, when making retrospective judgments, participants tended to exaggerate their experiences by recalling positive experiences (i.e., well-being) as more positive than they actually were and negative experiences (i.e., anxiety and stress) as more negative than they actually were.

Table 3 shows the results of the main analyses in the pooled sample. The table is organized as a 3 × 3 matrix such that the different item framings (Prompts 1–3) can be compared across columns, and the different time intervals (state, day, and week) can be compared across rows. We discuss these comparisons below. But first, we would like to point out that, overall, the baseline results in the first row replicate the finding that the retrospective well-being judgments were substantially but not perfectly aligned with the means of experience-sampled states ($r \approx .50$–$.80$ across samples). Further, across time intervals and item framings, 10 of the 18 peak/end effects that we tested were significant. These results suggest that, although these effects do not occur very consistently, they occur in many cases and are a likely explanation for misalignments between momentary and retrospective well-being judgments. The pattern of results was mostly robust when we excluded participants who had participated in more than one sample (Table A3). Including the temporal slope of well-being as a predictor did slightly reduce the effects of endpoints (as a result, the end effect only approached significance for Prompt 2) but did not significantly improve the model fit (Table A8).

Differences across item framings

Changing the framings of the retrospective judgment items influenced their associations with the mean of well-being states and the effects of the peaks and ends in the multiple regression models (comparing the *columns* in Table 3). In general, the mean across momentary well-being judgments was most strongly related to retrospective well-being judgments when participants were asked explicitly to estimate their average response in the state assessments (Prompt 3, see the first row of Table 3). The mean was less strongly related to the retrospective well-being judgments that referred to the assessment period (Prompt 2) than to the global well-being judgments (Prompt 1). Including the peak and end states in the prediction beyond the mean states

Table 2. Means, standard deviations, and intercorrelations for retrospective well-being judgments and momentary well-being characteristics in the pooled sample.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Retrospective WB (Prompt 1)	3.44	0.90									
2. Retrospective WB (Prompt 2)	4.18	1.09	**.63**								
3. Retrospective WB (Prompt 3)	4.26	1.02	**.74**								
4. Mean WB (State)	3.83	0.77	**.68**	**.58**	**.74**						
5. Peak WB (State)	5.49	0.66	**.48**	**.34**	**.42**	**.57**					
6. End WB (State)	3.86	1.24	**.51**	**.43**	**.50**	**.63**	**.37**				
7. Peak WB (Day)	5.11	0.74	**.58**	**.38**	**.61**	**.71**	**.78**	**.45**			
8. End WB (Day)	3.82	1.12	**.55**	**.44**	**.55**	**.71**	**.41**	**.90**	**.52**		
9. Peak WB (Week)	4.29	0.80	**.64**	**.51**	**.70**	**.89**	**.61**	**.59**	**.77**	**.67**	
10. End WB (Week)	3.83	0.92	**.63**	**.56**	**.64**	**.87**	**.49**	**.73**	**.62**	**.84**	**.81**

Note. N = 1,889. WB = well-being. Participants were assigned to either Prompt 2 or Prompt 3 in Samples 2, 3, and 4; therefore, no correlation could be calculated between the two prompts. Bold figures are significant at $p < .05$.

influenced the three retrospective judgments in diverging ways as well (see the second row of Table 3). Whereas the global well-being judgments in Prompt 1 were significantly related to both the peak and end states, only the end state was significantly related to the retrospective judgments in Prompt 2. By contrast, there were no significant effects of the peak and end states on the retrospective judgments in Prompt 3.

Differences across time intervals

Changing the time intervals to derive the peaks and ends of the state data also influenced the results (comparing the *rows* in Table 3). Generally speaking, the more broadly we operationalized the peaks and ends of the momentary well-being judgments, the stronger were their effects on the retrospective well-being judgments and the weaker was the effect of the mean level of state well-being. Moreover, the broader operationalization of the peaks and ends enhanced the diverging effects across the three item framings: Including the peak and end weeks in the model as predictors (see the last row of Table 3) resulted in significant peak and end effects for Prompt 1, significant end but not significant peak effects for Prompt 2, and significant peak but not significant end effects for Prompt 3. Comparing the models with versus without the peaks and ends showed that the models with peak and end *states* explained significantly more variance in the answers to Prompt 1, $F(2, 1885) = 43.35$, $p < .001$; models with peak and end *days* explained significantly more variance in the answers to Prompt 1, $F(2, 1885) = 50.37$, $p < .001$, and Prompt 3, $F(2, 370) = 4.83$, $p = .009$; and models with peak and end *weeks* explained significantly more variance in the answers to Prompt 1, $F(2, 1885) = 21.21$, $p < .001$, Prompt 2, $F(2, 407) = 4.16$, $p = .016$, and Prompt 3, $F(2, 370) = 5.73$, $p = .004$.

Differences across samples

Figure 1 visualizes the effects of mean, peak, and end well-being on the retrospective well-being judgments (the absolute standardized beta weights of each predictor in the regression analyses) for the pooled analyses and for the respective analyses in the single samples. Hence, this figure also provides an overview of the consistency of the effects across the four samples. As can be seen, even though there was some heterogeneity, the effects were moderately consistent across the samples. The detailed results of the analyses from the separate samples are presented in the supplement (Table A2).

Differences across state variables

The results of the specificity analyses on the association between momentary and retrospective judgments of other variables (anxiety, energy level, and stress) are presented in the supplement (Tables A5, A6, and A7). The pattern of effects was somewhat different for these variables. For example, we observed more pronounced peak than end effects for retrospective anxiety judgments, negative peak effects for retrospective energy-level judgments, and no end effects for retrospective stress judgments. Thus, peak/end effects in retrospective well-being judgments do not seem to be easy to generalize to retrospective judgments of other constructs.

Discussion

With the present research, we aimed to investigate the relevance of the peak-end rule for global and retrospective judgments of well-being in experience-sampled data. Systematic discrepancies between retrospective and momentary well-being judgments, especially those pertaining to the peak-end rule, have been well-documented in laboratory settings. However, findings on such effects in retrospective judgments of momentary well-being in everyday life have been less conclusive. By combining data from four ESM samples, comparing different item framings for retrospective judgments, and exploring time intervals of different lengths to identify peaks and ends in individuals' state data, we found evidence of systematic peak/end effects in retrospective well-being judgments in some but not all cases. Specifically, the effects were stronger for longer than shorter time intervals to identify the peaks and ends in individuals' state data, and they were stronger for global than more concrete retrospective judgments. These results suggest that researchers and practitioners should consider the possibility that retrospective judgments of everyday well-being disproportionately reflect individuals' peak and

Table 3. Predicting retrospective well-being judgments from the means, peaks, and ends of momentary well-being judgments across all samples.

	All samples (mega-analysis)											
	Retrospective WB—Prompt 1: *In general* (*N* = 1,889)				Retrospective WB—Prompt 2: *Over the past 4 weeks* (*N* = 411)				Retrospective WB—Prompt 3: *Average response* (*N* = 374)			
	b	95% CI	Test statistic	*p*	*b*	95% CI	Test statistic	*p*	*b*	95% CI	Test statistic	*p*
Mean WB (State)	**.676**	**[.643, .709]**	**39.83**	**< .001**	**.541**	**[.468, .614]**	**14.57**	**< .001**	**.707**	**[.642, .773]**	**21.29**	**< .001**
R^2	**.457**		**1586.66**	**< .001**	**.342**		**212.40**	**< .001**	**.549**		**453.32**	**< .001**
Mean WB (State)	**.515**	**[.468, .563]**	**21.50**	**< .001**	**.470**	**[.369, .571]**	**9.12**	**< .001**	**.673**	**[.580, .767]**	**14.19**	**< .001**
Peak WB (State)	**.133**	**[.093, .172]**	**6.60**	**< .001**	.016	[-.071, .102]	0.35	.723	.013	[-.076, .102]	0.29	.771
End WB (State)	**.136**	**[.094, .178]**	**6.39**	**< .001**	**.101**	**[.005, .196]**	**2.06**	**.040**	.044	[-.042, .131]	1.01	.312
R^2	**.481**		**581.53**	**< .001**	**.349**		**72.81**	**< .001**	**.551**		**151.15**	**< .001**
Mean WB (State)	**.442**	**[.385, .498]**	**15.33**	**< .001**	**.547**	**[.425, .669]**	**8.81**	**< .001**	**.577**	**[.464, .690]**	**10.05**	**< .001**
Peak WB (Day)	**.191**	**[.144, .237]**	**8.05**	**< .001**	−.068	[-.170, .035]	−1.30	.195	**.152**	**[.051, .252]**	**2.97**	**.003**
End WB (Day)	**.137**	**[.090, .183]**	**5.78**	**< .001**	.059	[-.045, .163]	1.11	.267	.035	[-.059, .128]	0.73	.468
R^2	**.484**		**590.14**	**< .001**	**.346**		**71.83**	**< .001**	**.561**		**157.43**	**< .001**
Mean WB (State)	**.411**	**[.325, .497]**	**9.33**	**< .001**	**.398**	**[.216, .580]**	**4.30**	**< .001**	**.494**	**[.351, .637]**	**6.79**	**< .001**
Peak WB (Week)	**.174**	**[.101, .248]**	**4.65**	**< .001**	−.038	[-.194, .119]	−0.47	.636	**.208**	**[.064, .352]**	**2.85**	**.005**
End WB (Week)	**.126**	**[.059, .194]**	**3.67**	**< .001**	**.202**	**[.064, .340]**	**2.88**	**.004**	.046	[-.069, .160]	0.78	.433
R^2	**.469**		**554.35**	**< .001**	**.355**		**74.66**	**< .001**	**.563**		**158.77**	**< .001**

Note. The table shows standardized beta weights (*b*) with 95% confidence intervals for the regression analyses in which momentary well-being characteristics (means, peaks, and ends) simultaneously predicted the respective retrospective well-being judgments as well as the percentage of variance explained by the respective model (R^2). Test statistics for beta weights are *t*-values; test statistics for R^2 are *F*-values. WB = well-being. Bold figures are significant at $p < .05$.

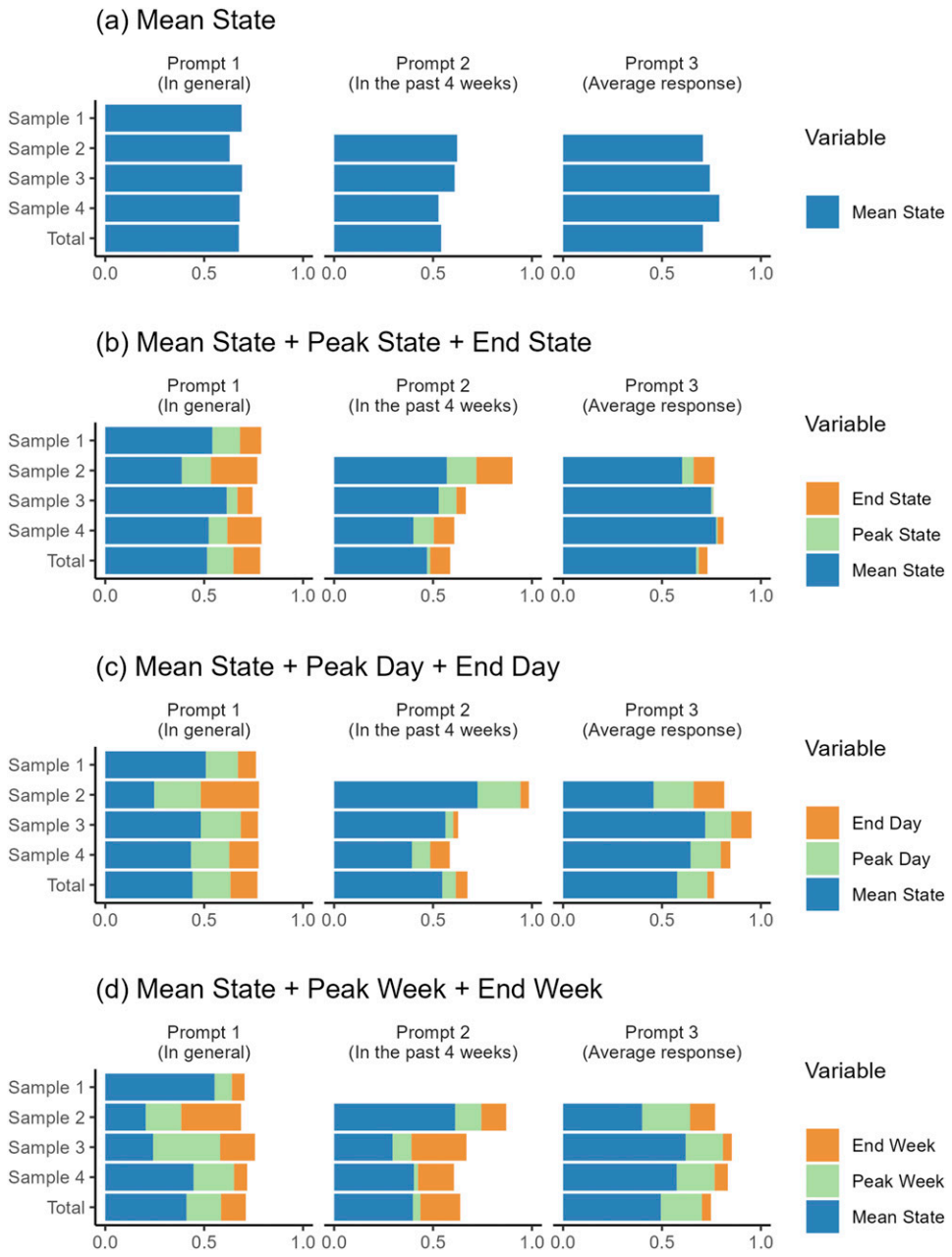

Figure 1. Visualization of effect sizes in the main analyses across samples.
Note. The figure shows the standardized beta weights (absolute values) of the respective predictors in the regression analyses for the pooled sample (bottom line in each panel) and for each sample separately. Panels (a)–(d) do not contain results for Prompts 2 and 3 in Sample 1, as these outcomes were not assessed in this sample. Note that absolute beta weights for the pooled analyses can appear smaller than the individual samples in some cases (e.g., see the effects for peak states in Prompt 2) because positive and negative beta weights in the regression analyses in the individual samples canceled each other out in the regression analyses in the pooled sample.

end experiences. At the same time, these influences were not severe enough to say that retrospective judgments are independent of momentary experiences or are "inaccurate" for that matter. Instead, even in the setups where retrospective diverged most strongly from momentary reports, they still reflected how individuals had felt on average to a substantial degree. Overall, our results point toward potential ways to mitigate discrepancies between retrospective and momentary judgments and arrive at more congruent retrospective estimates of momentary well-being. They also point to limitations of self-reports that need to be acknowledged in future study designs.

The peak-end rule for retrospective judgments of well-being in everyday life

Diverging effects of peak and end states, days, and weeks. Regarding the time interval that was used to determine the peaks and ends in individuals' state data, we generally found larger effects of peaks and ends for broader than narrower time intervals. For example, whereas including peak and end *states* as predictors of the retrospective well-being judgment next to the average well-being level significantly improved the fit of the model in only one of three cases, including peak and end *days* improved the fit in two out of three, and including peak and

end *weeks* improved the fit in all three cases. There are at least two potential statistical explanations and one potential theoretical explanation for this pattern of results.

On the statistical front, the broader conceptualizations of peaks and ends (daily/weekly level) have two "advantages" in the prediction of retrospective well-being judgments beyond mean-level effects compared with narrower conceptualizations (state level). First, the broader conceptualizations of peaks and ends can better differentiate between individuals. Whereas state peaks and ends can take on only an integer value between 1 and 6, the broader conceptualizations are aggregate scores of multiple assessments that can also result in noninteger values that express differences between individuals in a more nuanced fashion. Assessing momentary well-being on a more differentiated scale might reflect individual differences in experiences more validly and, as such, increase the effects of state peaks and ends compared with relying on single integer state values.

Second, estimations of broader conceptualizations share more data points with the estimation of mean momentary well-being. For example, for an individual who provided all of the possible 140 state assessments in Samples 3 or 4 (5 assessments per day for 28 days), the estimation of the peak *state* shares one data point, the estimation of the peak *day* shares five data points, and the estimation of the peak *week* shares 35 data points with the estimation of the overall mean. Accordingly, broader conceptualizations of peaks and ends share more variance with the mean (i.e., they are more strongly correlated with it), which results in stronger effects of the peaks and ends in the multiple regression analyses next to the mean than the narrower conceptualization. These statistical artifacts should be kept in mind when interpreting the stronger peak-end effects for the broader time intervals.

However, an analogous theoretical explanation is also possible. That is, in forming an estimate of their well-being, individuals might frame the memory of a single positive moment (i.e., the peak state) as less indicative of their overall well-being compared with a memory of a great day (i.e., the peak day) or a glorious week (i.e., the peak week). Thus, individuals might assign more weight to particularly positive days or weeks than single moments when retrospectively judging their momentary well-being over a certain period of time. This theoretical explanation is related to the statistical one above: Good moments are less indicative of a good month (i.e., the entire ESM period) than good days/weeks are because the moments make up a smaller portion of that period.

Diverging effects on different kinds of retrospective judgments. Besides depending on the applied time interval, there were diverging peak and end effects that depended on the framing of the retrospective well-being judgment items, and these effects were magnified as the time intervals became broader. For the global retrospective well-being judgments in Prompt 1 ("*In general, I feel happy*"), we found significant effects of peaks and ends beyond mean-level effects for all time intervals. For the retrospective judgment in Prompt 2 ("*Over the past 4 weeks, I felt happy*"), we found only significant end effects that were weaker and nonsignificant for the end day but stronger and significant for the end week. By contrast, for the retrospective judgment in Prompt 3 ("*Over the past 4 weeks, you have regularly indicated how you felt. What would you say was your average response to the statement 'I feel happy*'?"), we found no significant effects of peak and end states, but we found significant effects of peak days and peak weeks.

In the Introduction, we argued that different predictions can be derived about the influence of item framings on the basis of either dual-process theories of cognition (e.g., Evans, 2003; Evans & Stanovich, 2013) or the accessibility model of emotional self-report (Robinson & Clore, 2002). The pattern of results obtained in this study might be interpreted as reflecting a combination of the two perspectives. According to dual-process theories, more familiar and unconscious tasks should elicit fast, intuitive (System 1) cognitive processes, whereas tasks that are novel or require deliberate thought should elicit more elaborate, analytical (System 2) cognitive processes. Regarding the three item framings, Prompt 1 represented a standard, global statement about oneself ("In general…") that is comparable to items most frequently applied in classic trait questionnaires. As such, it likely entices participants to rely on default System 1 operations to arrive at a quick estimate of their overall well-being. These operations represent a useful and to some degree "accurate" shortcut for retrospectively evaluating one's momentary well-being, as can be seen in the overall high correlation between the retrospective judgments in Prompt 1 and the mean of state well-being, even after peaks and ends were included in the prediction. However, they are also more prone to rely on heuristics, as can be seen in the disproportionate effects of the peaks and ends beyond the mean of the experience-sampled data on the retrospective well-being judgments.

By contrast, Prompt 2 referred to a more specific episode ("Over the past 4 weeks…") that might entice participants to recall episodic memories about the period and try to integrate all this information in an elaborate, System 2 fashion. This approach might reduce the likelihood that single experiences (e.g., peak experiences) will have a disproportionate influence on the retrospective judgment. However, it does not represent a viable approach for arriving at "accurate" retrospective judgments (i.e., retrospective judgments that are closely related to the mean of the experience-sampled well-being data) because participants might overly emphasize some experiences over others due to the cognitively demanding nature of the integration process. As a consequence of being overwhelmed by the challenge of forming a comprehensive retrospective judgment, participants might default to simpler heuristics, such as how they are feeling at the moment, which might explain the significant recency (i.e., end) effects (e.g., Brose et al., 2013; Eich et al., 1985). Overall, this resulted in retrospective judgments that were less closely related to the actual (average) level of experienced well-being compared with the other two item framings.

Finally, Prompt 3 might also encourage more elaborate System 2 operations. However, in contrast to Prompt 2, it provides participants support in integrating the episodic information because it encourages them not to think about

all their experiences in the period but to refer to the specific moments in which they provided the momentary well-being assessments ("What was your average response…"). This way, participants successfully form elaborate, comprehensive retrospective judgments of their previously reported momentary well-being, resulting in the highest association between retrospective and average momentary well-being judgments for Prompt 3. Furthermore, these retrospective judgments seem to be less strongly influenced by peak and end states. Of note, this reasoning does not apply to the same degree to peak days and weeks, which seem to disproportionally influence these specific retrospective well-being judgments as well.

In a nutshell, Prompts 2 and 3 might encourage a more elaborate System 2 operation, but this approach is successful only if the episodic information is accessible and succinct enough to be integrated across the period in question. Whereas such processes might explain the pattern of results obtained here, it is important to emphasize the exploratory nature of our study. Future studies might deduce concrete hypotheses from these considerations and test them with a confirmatory approach to enhance our understanding of the effects of item framings on the accuracy of retrospective well-being judgments. Rather than *explaining* the diverging effects of the item framings, our study shows that these diverging effects *exist*, indicating that there are potential ways to mitigate the discrepancies between retrospective and momentary well-being judgments and arrive at more accurate estimates of individuals' aggregated momentary well-being when one lacks the time and resources to assess it directly via ESM or comparable methods. For example, providing participants support in forming elaborate, comprehensive judgments by referring to concrete situations (in this case, answering the brief ESM surveys) could be one tool to increase the accuracy of retrospective judgments. This approach already gets applied in the clinical context with the timeline follow back method (e.g., to assist recovering alcoholics in retrospectively recalling alcohol consumption, Sobell & Sobell, 1992) and in the day reconstruction method (Kahneman et al., 2004) to increase the accuracy of retrospective end-of-day well-being judgments. Our data suggest that it might also be effective for retrospective judgments over longer periods of time.

Momentary, retrospective, and global judgments: Different selves?

In the Introduction, we discussed that some researchers have asserted that retrospective judgments are "biased" on the basis of their relationship with momentary assessments. However, it should be noted that not all researchers will agree to that position. As Conner and Barrett (2012) pointed out, recent perspectives have moved from regarding retrospective judgments as a biased, less valid approach for approximating momentary judgments to conceptualizing each as capturing a distinct aspect of the self (e.g., Kahneman & Riis, 2007): The *experiencing self* can be measured only by using momentary ratings, as it relies on experiential information from that moment (e.g., physiological states and perceptual input from a particular time and context) that cannot be accessed after the episode has passed. On the contrary, retrospective and global reports capture the *remembering self*, which relies on episodic information (i.e., reconstructive processes and the long-term memory system), and the *believing self*, which emphasizes semantic information (i.e., one's self-concept).

In keeping with that theme, the disproportional effects of peaks and ends particularly for the global well-being judgments might be regarded as a reasonable inference as part of a schematic self-cognitive process (i.e., the formation of the remembering and believing self). Just as participants rely on peak and end days/weeks stronger because they are more indicative of their experience in the past month than single states, peaks and ends might per se be more relevant for the formation of self-related schemata than other experiences. On the one hand, peak experiences might carry information about situations that rarely occur but that are especially relevant to one's self-concept. For example, individuals might be less able to judge their level of Neuroticism on the basis of the many mundane experiences during a month, whereas this judgment might be well informed by a handful of stressful situations in the same period. On the other hand, end experiences might be particularly relevant for schematic processing because they carry information about how one's overall life situation has developed in the past month. Judging such trajectories in well-being necessarily requires information about one's recent experiences, which is why it makes sense that these influence individuals' self-schemata more strongly than experiences longer ago. As such, peak/end effects might be regarded not as a bias but as a reasonable inference mechanism.

In our view, this perspective emphasizes an important point to consider when contemplating the relationships between momentary, retrospective, and global judgments of well-being: None of the three approaches is necessarily superior to the others, and all come with their unique strengths. Momentary judgments are the best way to directly assess participants' actual experience (i.e., the experiencing self), and experience-sampled state data offer other advantages, such as the opportunity to investigate within-person variability (e.g., Baird et al., 2006; Geukes et al., 2017; Scharbert, Dein, et al., 2024), within-person contingencies (Kroencke et al., 2023; e.g., Kuper et al., 2022), or state trajectories over time (Buecker et al., 2020; Scharbert, Humberg, et al., 2024). By contrast, retrospective and global reports have been found to be superior in predicting decision-based outcomes (e.g., Conner & Barrett, 2012; Wirtz et al., 2003). The aspect of the self that researchers want to investigate depends on their specific objective, and which aspect one emphasizes in one's own life arguably comes down to personal preference. However, our findings underscore the importance of at least conceptually distinguishing between these three aspects of the self to ensure valid inferences for one or the others.

Limitations and future directions

Our study relied on four large ESM data sets, investigated peak and end effects that went beyond mean-level effects, and linked retrospective judgments to momentary rather

than daily well-being judgments to address the inconsistencies and methodological limitations of prior studies on peak and end effects in retrospective well-being judgments. Nevertheless, it also comes with limitations that provide directions for future research.

First, the assessment of momentary well-being might be improved in future studies. Here, we relied on only one momentary well-being item that was assessed on a rather coarse 6-point scale per time point. A well-being measure based on several items that involve a more differentiated response format (e.g., ranging from 1 to 100) would have been desirable to obtain more reliable assessments and to better reflect individual differences in state well-being. Moreover, whereas we measured state well-being directly in the moment in Samples 3 and 4, we measured state well-being only approximately in the moment in Samples 1 and 2, as the ESM surveys asked participants to refer to their last social interaction or individual activity in the past hour in these samples. Although the last interaction/activity was arguably very close to the time when the participants were completing the ESM survey, we cannot rule out the possibility that some form of processing of the experiences had already taken place. Even though we did not find systematic differences in the results between Samples 1 and 2 and Samples 3 and 4, future studies should strive for a uniform and more direct assessment of state well-being in the moment.

Second, our samples comprised a majority of young and female participants from Western, educated, industrialized, rich, and democratic (WEIRD; Henrich et al., 2010) countries. This restricted demography in our samples might reduce the generalizability of the results presented here. For example, Ready et al. (2007) found diverging patterns of recall biases for participants of different ages, as older participants tended to retrospectively overestimate positive affect more strongly than younger participants, who in turn tended to retrospectively overestimate negative affect more strongly than older participants. Such diverging effects might also be found for other demographic characteristics, a point that should be kept in mind when considering the results reported here.

Third, as a direction for future research, we deem it important to systematically consider the diverging peak and end effects that we found for constructs closely related to individuals' well-being (i.e., retrospective judgments of anxiety, energy level, and stress). These findings suggest that peak/end effects in retrospective well-being judgments cannot easily be generalized to related constructs, and future studies should target and compare these retrospective judgments. This observation is in line with the results from previous studies in which diverging peak and end effects for positive versus negative affect were identified (e.g., Ganzach & Yaor, 2019; Parkinson et al., 1995).

Fourth, besides identifying peak/end effects, we also found evidence of other systematic discrepancies between retrospective and momentary judgments that have been reported in the literature before. Specifically, we found that individuals retrospectively exaggerated their well-being levels, as, on average, retrospective well-being judgments were significantly more positive than the actual mean of momentary well-being from ESM ratings. However, this effect also occurred for the relationship between the retrospective and momentary anxiety and stress judgments, in that the retrospective anxiety and stress judgments were significantly higher than the actual mean levels of experienced anxiety and stress. This finding is in line with prior research, which showed a general *extremity bias* in which individuals recalled positive experiences as more positive and negative experiences as more negative than they actually were (e.g., Ben-Zeev et al., 2009; Ellison et al., 2020). Investigating how these effects are related to and interact with those pertaining to the peak-end rule might be a fruitful avenue for future research.

Conclusion

According to the peak-end rule, retrospective well-being judgments are disproportionately influenced by the peak and the end of an episodic experience. Whereas experimental studies in laboratory settings have provided empirical support for this notion, studies on data obtained outside the laboratory have been less conclusive. Across four ESM samples, we found disproportionate peak and end effects that went beyond the mean level effects on retrospective well-being judgments. Put differently, even though retrospective reports reflected how individuals felt to a substantial degree on average, they overemphasized individuals' peak and end experiences. However, these effects depended on the framing of the retrospective judgment item (global vs. more specific framings) and broad versus narrow conceptualizations of peaks and ends (states, days, and weeks). These findings thus suggest potential ways to mitigate peak/end effects in retrospective well-being judgments, for example, by referring to concrete situations when asking for retrospective evaluations. Overall, our results highlight the importance of conceptually distinguishing between momentary and retrospective well-being judgments and of intentionally choosing the most appropriate measurement approach on the basis of such conceptual considerations.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This research was supported by funding provided to Mitja D. Back, Markus Bühner, and Maarten van Zalk by the German Research Foundation under grant numbers BA 3731/11-1, BU 1414/3-1, and ZA 1147/1-1. The funders played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are grateful to Carolin Albrecht, Cosima Heinen, and Anna Schnickmann for their help conducting and documenting the data collection or searching the literature.

Open science statement

The study materials, data, and analysis scripts used for this article can be accessed at https://osf.io/32b8a/. We embrace the values of openness and transparency in science (Schönbrodt

et al., 2015). Therefore, we follow the 21-word solution (Simmons et al., 2012) or refer to project documentation on the OSF. The data and statistical code necessary to reproduce the reported results can be retrieved from https://osf.io/32b8a/. The hypotheses for the study, data cleaning procedures, and statistical analyses were preregistered; see https://osf.io/32b8a/.

ORCID iDs

Julian Scharbert https://orcid.org/0000-0003-3020-2976
Katharina Utesch https://orcid.org/0000-0002-7424-306X
Thomas Reiter https://orcid.org/0000-0002-7443-4173
Julian ter Horst https://orcid.org/0000-0002-2555-4742
Maarten van Zalk https://orcid.org/0000-0002-0185-8805
Mitja D Back https://orcid.org/0000-0003-2186-1558
Richard Rau https://orcid.org/0000-0002-1490-8584

Supplemental Material

Supplemental material for this article is available online.

Notes

1. See Horwitz et al. (2023) for a recent investigation of retrospective biases in the PHQ-9, which assesses the frequency of depressive symptoms during the 2 weeks prior to the assessment.
2. The fifth brief survey on each day was sent out at a fixed time each day at 8 pm rather than at a random time point in Samples 3 and 4.
3. Due to a technical error in the smartphone research app, no data on momentary stress judgments were collected in Sample 4 for participants who used the smartphone research app. However, these data were collected for the SoSci Survey participants, who made up the majority of participants in that sample.
4. Besides conceptualizing the peak as the most positive (highest) value, one might also conceptualize the peak as a participant's most extreme (highest or lowest) value. As preregistered, we considered both conceptualizations but focused our interpretation on the conceptualization in which the peak contributed more to explaining variance in the retrospective judgments. The "extreme" peak was retrieved by *z*-standardizing all assessments within persons and identifying the highest absolute score. The peak of the state data was the raw (unstandardized) value of this measurement occasion. If a participant had two peaks (i.e., if the highest and lowest values for that participant had the same absolute *z*-score), we assigned the person mean as the score for the peak point so as not to influence the estimation of the slope of the peak in the multiple regression analyses. Comparing the "high" peak with the "extreme" peak consistently showed stronger effects of the "high" peak on the retrospective well-being judgments than the "extreme" peak, and multiple regression models including the "high" peak explained more variance in the retrospective judgments than those including the "extreme" peak in seven out of nine cases. Therefore, as preregistered, we focused on the models with the "high" peak (henceforth simply referred to as the "peak") in this manuscript, and we report the results of the models with the "extreme" peak in the supplement (see Table A4).
5. Prompt 1 was assessed on a different scale than the momentary judgments (5-point vs. 6-point Likert scale). Thus, the two cannot be compared directly.

References

Alaybek, B., Dalal, R. S., Fyffe, S., Aitken, J. A., Zhou, Y., Qu, X., Roman, A., & Baines, J. I. (2022). All's well that ends (and peaks) well? A meta-analysis of the peak-end rule and duration neglect. *Organizational Behavior and Human Decision Processes*, *170*, 104149. https://doi.org/10.1016/j.obhdp.2022.104149.

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). Formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods*, *52*(1), 376–387. https://doi.org/10.3758/s13428-019-01236-y

Baird, B. M., Le, K., & Lucas, R. E. (2006). On the nature of intraindividual personality variability: Reliability, validity, and associations with well-being. *Journal of Personality and Social Psychology*, *90*(3), 512–527. https://doi.org/10.1037/0022-3514.90.3.512

Ben-Zeev, D., Young, M. A., & Madsen, J. W. (2009). Retrospective recall of affect in clinically depressed individuals and controls. *Cognition and Emotion*, *23*(5), 1021–1040. https://doi.org/10.1080/02699930802607937

Brose, A., Lindenberger, U., & Schmiedek, F. (2013). Affective states contribute to trait reports of affective well-being. *Emotion*, *13*(5), 940–948. https://doi.org/10.1037/a0032401

Buecker, S., Horstmann, K. T., Krasko, J., Kritzler, S., Terwiel, S., Kaiser, T., & Luhmann, M. (2020). Changes in daily loneliness for German residents during the first four weeks of the COVID-19 pandemic. *Social Science and Medicine*, *265*, 113541. https://doi.org/10.1016/j.socscimed.2020.113541.

Cohen, S., Doyle, W. J., Skoner, D. P., Fireman, P., Gwaltney, J. M., & Newsom, J. T. (1995). State and trait negative affect as predictors of objective and subjective symptoms of respiratory viral infections. *Journal of Personality and Social Psychology*, *68*(1), 159–169. https://doi.org/10.1037//0022-3514.68.1.159

Conner, T. S., & Barrett, L. F. (2012). Trends in ambulatory self-report: The role of momentary experience in psychosomatic medicine. *Psychosomatic Medicine*, *74*(4), 327–337. https://doi.org/10.1097/PSY.0b013e3182546f18

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of Nervous and Mental Disease*, *175*(9), 526–536. https://doi.org/10.1097/00005053-198709000-00004

Diener, E., Wirtz, D., & Oishi, S. (2001). End effects of rated life quality: The James Dean Effect. *Psychological Science*, *12*(2), 124–128. https://doi.org/10.1111/1467-9280.00321

Eich, E., Reeves, J. L., Jaeger, B., & Graff-Radford, S. B. (1985). Memory for pain: Relation between past and present pain intensity. *Pain*, *23*(4), 375–380. https://doi.org/10.1016/0304-3959(85)90007-7

Eisenhauer, J. G. (2021). Meta-analysis and mega-analysis: A simple introduction. *Teaching Statistics*, *43*(1), 21–27. https://doi.org/10.1111/test.12242

Ellison, W. D., Trahan, A. C., Pinzon, J. C., Gillespie, M. E., Simmons, L. M., & King, K. Y. (2020). For whom, and for what, is experience sampling more accurate than retrospective report? *Personality and Individual Differences*, *163*, 110071. https://doi.org/10.1016/j.paid.2020.110071.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The Obsessive-Compulsive Inventory: Development and validation of a short version. *Psychological Assessment*, *14*(4), 485–496. https://doi.org/10.1037/1040-3590.14.4.485

Fredrickson, B. L. (2000). Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition and Emotion*, *14*(4), 577–606. https://doi.org/10.1080/026999300402808

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65*(1), 45–55. https://doi.org/10.1037/0022-3514.65.1.45

Ganzach, Y., & Yaor, E. (2019). The retrospective evaluation of positive and negative affect. *Personality and Social Psychology Bulletin*, *45*(1), 93–104. https://doi.org/10.1177/0146167218780695

Geng, X., Chen, Z., Lam, W., & Zheng, Q. (2013). Hedonic evaluation over short and long retention intervals: The mechanism of the peak-end rule. *Journal of Behavioral Decision Making*, *26*(3), 225–236. https://doi.org/10.1002/bdm.1755

Geukes, K., Nestler, S., Hutteman, R., Dufner, M., Küfner, A. C. P., Egloff, B., Denissen, J. J. A., & Back, M. D. (2017). Puffed-up but shaky selves: State self-esteem level and variability in narcissists. *Journal of Personality and Social Psychology*, *112*(5), 769–786. https://doi.org/10.1037/pspp0000093

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hessel, A., Schumacher, J., Geyer, M., & Brähler, E. (2001). Symptom-Checkliste SCL-90-R: Testtheoretische Überprüfung und Normierung [Symptom-Checklist SCL-90-R: Test theoretical validation and standardization]. *Diagnostica*, *47*(1), 27–39. https://doi.org/10.1026//0012-1924.47.1.27

Horwitz, A. G., Zhao, Z., & Sen, S. (2023). Peak-end bias in retrospective recall of depressive symptoms on the PHQ-9. *Psychological Assessment*, *35*(4), 378–381. https://doi.org/10.1037/pas0001219

Jiang, L., Yin, D., & Liu, D. (2019). Can joy buy you money? The impact of the strength, duration, and phases of an entrepreneur's peak displayed joy on funding performance. *Academy of Management Journal*, *62*(6), 1848–1871. https://doi.org/10.5465/amj.2017.1423

Kahneman, D. (2011). *Thinking, fast and slow*. Penguin.

Kahneman, D., & Frederick, S. (2012). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases* (pp. 49–81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*(6), 401–405. https://doi.org/10.1111/j.1467-9280.1993.tb00589.x

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, *306*(5702), 1776–1780. https://doi.org/10.1126/science.1103572

Kahneman, D., & Riis, J. (2007). Living, and thinking about it: Two perspectives on life. In F. A. Huppert (Ed.), *The science of well-being reprinted* (pp. 284–305). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198567523.003.0011

Kam, C. C. S. (2018). Why do we still have an impoverished understanding of the item wording effect? An empirical examination. *Sociological Methods & Research*, *47*(3), 574–597. https://doi.org/10.1177/0049124115626177

Kritzler, S., Krasko, J., & Luhmann, M. (2020). Inside the happy personality: Personality states, situation experience, and state affect mediate the relation between personality and affect. *Journal of Research in Personality*, *85*, 103929. https://doi.org/10.1016/j.jrp.2020.103929.

Kroencke, L., Kuper, N., Mota, S., Geukes, K., Zeigler-Hill, V., & Back, M. D. (2023). Narcissistic status pursuit in everyday social life: A within-person process approach to the behavioral and emotional dynamics of narcissism. *Journal of Personality and Social Psychology*, *125*(6), 1519–1541. https://doi.org/10.1037/pspp0000467

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kuper, N., Breil, S. M., Horstmann, K. T., Roemer, L., Lischetzke, T., Sherman, R. A., Back, M. D., Denissen, J. J. A., & Rauthmann, J. F. (2022). Individual differences in contingencies between situation characteristics and personality states. *Journal of Personality and Social Psychology*, *123*(5), 1166–1198. https://doi.org/10.1037/pspp0000435

Leiner, D. J. (2019). *SoSci survey*. https://www.soscisurvey.de

Neubauer, A. B., Scott, S. B., Sliwinski, M. J., & Smyth, J. M. (2020). How was your day? Convergence of aggregated momentary and retrospective end-of-day affect ratings across the adult life span. *Journal of Personality and Social Psychology*, *119*(1), 185–203. https://doi.org/10.1037/pspp0000248

Parkinson, B., Briner, R. B., Reynolds, S., & Totterdell, P. (1995). Time frames for mood: Relations between momentary and generalized ratings of affect. *Personality and Social Psychology Bulletin*, *21*(4), 331–339. https://doi.org/10.1177/0146167295214003

R Core Team. (2022). *R: A language and environment for statistical computing [computer software]*. R Foundation for Statistical Computing. https://www.R-project.org/

Ready, R. E., Weinberger, M. I., & Jones, K. M. (2007). How happy have you felt lately? Two diary studies of emotion recall in older and younger adults. *Cognition & Emotion*, *21*(4), 728–757. https://doi.org/10.1080/02699930600948269

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, *66*(1), 3–8. https://doi.org/10.1016/0304-3959(96)02994-6

Redelmeier, D. A., Katz, J., & Kahneman, D. (2003). Memories of colonoscopy: A randomized trial. *Pain*, *104*(1-2), 187–194. https://doi.org/10.1016/s0304-3959(03)00003-4

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report.

Psychological Bulletin, *128*(6), 934–960. https://doi.org/10.1037/0033-2909.128.6.934

Röcke, C., Hoppmann, C. A., & Klumb, P. L. (2011). Correspondence between retrospective and momentary ratings of positive and negative affect in old age: Findings from a one-year measurement burst design. *The Journals of Gerontology*, *66*(4), 411–415. https://doi.org/10.1093/geronb/gbr024

Scharbert, J., Dein, L. M., Kroencke, L., Nestler, S., Back, M. D., & Utesch, K. (2024). Narcissists' affective well-being: Associations of grandiose narcissism with state affect level and variability. *Journal of Personality and Social Psychology*. Advance online publication. https://doi.org/10.1037/pspp0000495

Scharbert, J., Humberg, S., Kroencke, L., Reiter, T., Sakel, S., ter Horst, J., Utesch, K., Gosling, S., Harari, G., Matz, S., Schoedel, R., Stachl, C., Aguilar, N. M. A., Amante, D., Aquino, S. D., Bastias, F., Bornamanesh, A., Bracegirdle, C., Chauvin, B., … Back, M. D. (2024). Psychological well-being in Europe after the outbreak of war in Ukraine. *Nature Communications*, *15*(1), 1202. https://doi.org/10.1038/s41467-024-44693-6

Scharbert, J., Reiter, T., Sakel, S., ter Horst, J., Geukes, K., Gosling, S. D., Harari, G. M., Kroencke, L., Matz, S., Schoedel, R., Shani, M., Stachl, C., Talaifar, S., Aguilar, N. M. A., Amante, D., Aquino, S. D., Bastias, F., Biesanz, J. C., Bornamanesh, A., … Back, M. D. (2023). A global experience-sampling method study of well-being during times of crisis: The CoCo project. *Social and Personality Psychology Compass*, *17*(10), Article e12813. https://doi.org/10.1111/spc3.12813

Schönbrodt, F. D., Maier, M., Heene, M., & Zehetleitner, M. (2015). *Commitment to research transparency*. https://www.researchtransparency.org

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105. https://doi.org/10.1037/0003-066X.54.2.93

Schwarz, N. (2012). Why researchers should think "real-time": A cognitive rationale. In M. R. Mehl, & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 22–42). Guilford.

Scollon, C. N., Howard, A. H., Caldwell, A. E., & Ito, S. (2009). The role of ideal affect in the experience and memory of emotions. *Journal of Happiness Studies*, *10*(3), 257–269. https://doi.org/10.1007/s10902-007-9079-9

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21-word solution*. https://doi.org/10.2139/ssrn.2160588

Sobell, L., & Sobell, M. (1992). Timeline follow-back: A technique for assessing self-reported alcohol consumption. In R. Z. Litten, & J. P. Allen (Eds.), *Measuring alcohol consumption: Psychosocial and biological methods* (pp. 41–72). Humana Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Tivian. (2022). *EFS survey*. Tivian XI GmbH. https://www.unipark.de/

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wilhelm, P., Perrez, M., & Pawlik, K. (2012). Conducting research in daily life: A historical review. In M. R. Mehl, & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 62–86). Guilford.

Wilt, J., Noftle, E. E., Fleeson, W., & Spain, J. S. (2012). The dynamic role of personality states in mediating the relationship between extraversion and positive affect. *Journal of Personality*, *80*(5), 1205–1236. https://doi.org/10.1111/j.1467-6494.2011.00756.x

Wirtz, D., Kruger, J., Napa Scollon, C., & Diener, E. (2003). What to do on spring break? The role of predicted, on-line, and remembered experience in future choice. *Psychological Science*, *14*(5), 520–524. https://doi.org/10.1111/1467-9280.03455

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, *30*(3), 825–846. https://doi.org/10.1177/10731911211067538

Zygar-Hoffmann, C., & Schönbrodt, F. D. (2020). Recalling experiences: Looking at momentary, retrospective and global assessments of relationship satisfaction. *Collabra: Psychology*, *6*(1). 7.https://doi.org/10.1525/collabra.278