UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Zygar-Hoffmann, C., & Schönbrodt, F. D. (2020). Recalling Experiences: Looking at Momentary, Retrospective and Global Assessments of Relationship Satisfaction. *Collabra: Psychology,* 6(1): 7. DOI: https://doi.org/10.1525/collabra.278

ORIGINAL RESEARCH REPORT

# Recalling Experiences: Looking at Momentary, Retrospective and Global Assessments of Relationship Satisfaction

Caroline Zygar-Hoffmann and Felix D. Schönbrodt

Relationship satisfaction can be assessed in retrospection, as a global evaluation, or as a momentary state. In two experience sampling studies ($N$ = 130, $N$ = 510) the specificities of these assessment modalities are examined. We show that 1) compared to other summary statistics like the median, the mean of relationship satisfaction states describes retrospective and global evaluations best (but the difference to some other summary statistics was negligible); 2) retrospection introduces an overestimation of the average annoyance in the relationship reported on a momentary basis, which results in an overall negative mean-level bias for retrospective relationship satisfaction; 3) this bias is most strongly moderated by global relationship satisfaction at the time of retrospection; 4) snapshots of momentary relationship satisfaction get representative of global evaluations after approximately two weeks of sampling. The findings extend the recall bias reported in the literature for retrospection of negative affect to the domain of relationship evaluations and assist researchers in designing efficient experience sampling studies.

**Keywords:** romantic relationships; relationship satisfaction; retrospection; recall; experience sampling method

Global evaluations of individuals' experiences should correspond to their daily experiences. Fleeson (2001) elaborated on this relationship between global evaluations and momentary behavior in the personality domain and described personality traits as density distributions of personality states. The reasoning that traits reflect to some degree characteristics of the occurrence of corresponding states (such as the amount or intensity) is also common for other psychological constructs, such as affective traits, mood and emotions (Rosenberg, 1998).

States are often assumed to be dynamic and affected by situational influences and must therefore be assessed in the moment, for example with the experience sampling method (ESM; Csikszentmihalyi & Larson, 1987). Traits on the other hand are most commonly conceptualized as stable dispositions, typically assessed with self-reports of individuals' global representations of their behaviors and experiences. These trait evaluations have much in common with a third assessment mode: The summative recall of experiences during a certain time period, also called retrospective assessment (e.g., used for the Positive and Negative Affect Schedule, which asks individuals to evaluate their affect during the last day(s), week(s), month(s) or year(s), Watson, Clark, & Tellegen, 1988).

Retrospective assessments can introduce recall biases: For instance, studies find discrepancies between individuals' recall of affective experiences and their momentary report in ESM during that time. More global (trait) evaluations are prone to similar biases as well, as they require to appraise an even wider and more unspecific range of situations and time (Baumert et al., 2017; Reis & Gable, 2000; Robinson & Clore, 2002b).

As a lot of emotional experiences happen within relationships, we explore the correspondence between individuals' state assessments of their relationship satisfaction, measured repeatedly with ESM, and their global as well as retrospective assessment of their relationship satisfaction in two studies. Our aim is to inform researchers about (1) the way ESM data on relationship satisfaction relates to classical measurement tools, by investigating to what extent the average, most intense, or more recent experience corresponds to retrospection and global assessments; (2) the differential validity of retrospective assessments, by investigating what kind of bias in retrospection occurs; (3) the role individual differences have in recalling the past, by investigating the moderation of recall biases by traits, global relationship satisfaction, and other individual or relationship characteristics; (4) the optimal design of ESM studies with high accuracy, by investigating what level of aggregation is sufficient to approach a reliable measurement of the global index.

## The Special Case of Relationship Satisfaction

Our study focused on a dyadic setting and the assessments of individuals' relationship satisfaction. While this construct naturally plays a vital role for the study of relationships, it is also of special interest from an assessment perspective.

Department of Psychology, Ludwig-Maximilians-Universität München, DE

Corresponding author: Caroline Zygar-Hoffmann (caroline.zygar@psy.lmu.de)

On the one hand, the affective component of relationship satisfaction allows for a comparison with the study of concrete affective experiences, like pain or specific emotions. On the other hand, the construct has trait-like features: It reflects an inter-individual difference, is mainly assessed by asking individuals to globally evaluate their feelings, behavior and experiences (with regard to their relationships; Fincham & Rogge, 2010) and is related to the average of correspondent everyday states (e.g., Hofmann, Finkel, & Fitzsimons, 2015; Zygar et al., 2018a). Furthermore, global relationship satisfaction typically shows medium to strong stability in couples that do not break up (e.g., $r$ = .61–.69 over two years, which is close to typical personality trait stabilities across the same period of time, Fallis, Rehman, Woody, & Purdon, 2016; McCrae, Bond, Yik, Trapnell, & Paulhus, 1998). Studying the assessment of relationship satisfaction can therefore not only contribute to the understanding of this specific construct, it can also provide insights that might be relevant for the related literature on biases occurring during the assessment of affective experiences and traits more generally.

### What Summary Statistic of States Corresponds Best to Retrospection and Global Assessments? (RQ1)

Our first goal was to examine the way ESM assessments relate to classical measurement tools. The distribution of an individual's momentary feelings or behaviors can be summarized across different time periods by various measures, such as the central tendency or extreme values. Which measure best represents what individuals do when they retrospectively assess a time period or globally evaluate their relationship?

For the recall of daily mood, studies found that the peak mood describes retrospection better than or incremental to the average mood (Hedges, Jandorf, & Stone, 1985; Parkinson, Briner, Reynolds, & Totterdell, 1995). This is in line with findings from personality, showing that while the average of personality states is the best indicator for global trait measures, the maximum of the state experience is incrementally relevant (Fleeson & Gallagher, 2009). For the recall of pain and various affective experiences during single, discrete events, a series of studies found that not only the most intense, but also the most recent events are predominant for the evaluation of the experience, termed the peak-and-end rule (see Fredrickson, 2000 for a review). However, this rule seems to have only limited value for multi-episodic events like days, where longer time periods are considered, which are characterized by a mix of events and emotions (Miron-Shatz, 2009).

In sum, previous research found evidence for the informational value of averages, peaks and recent experiences. For relationship satisfaction, we a priori did not have a hypothesis about what summary statistic best describes the retrospective and global assessment. We therefore examined the central tendency (mean and median), extreme values (90% and 10% quantile), and recency effects (mean during the last week and the last day of the ESM period), contrasted with a primacy effect (mean during the first week).

### What Bias Occurs in Retrospection? (RQ2)

Our second goal was to investigate whether individuals are biased in their retrospective assessment of their relationship satisfaction. When it comes to evaluating the convergence of judgments, it is possible to differentiate at least two aspects (see e.g., Fletcher & Kerr, 2010; Neubauer, Scott, Sliwinski, & Smyth, 2019; West & Kenny, 2011): First, mean-level bias (also called directional bias or level convergence), which refers to the sample mean of a judgment being different from the sample mean of another judgment that is used as an external reference category (i.e., as truth criterion). In our case, the external reference is a certain summary of an individual's own repeated assessment of relationship satisfaction with ESM, which is compared to that individual's retrospective assessment. A second aspect that can be considered is tracking accuracy (also called truth force or correspondence convergence), which refers to the actual relationship between the reference category (or truth criterion) and the judgments. In our studies, we investigate tracking accuracy in form of the between-person effect of the aggregated ESM assessments on individuals' retrospective judgments.

In this reasoning, discrepancies between retrospection and mean of ESM states are regarded as systematic recall errors caused during retrospection. However, as already pointed out by others (e.g., Conner & Feldman Barrett, 2012; Feldman Barrett, 1997), it may be that retrospective evaluations are in fact more accurate or have higher validity in some contexts, also because they target all experiences during the examined period, even those moments that were not captured by the ESM surveys. It seems to depend on the type of construct and the type of prediction, whether aggregated ESM states, retrospection or global self-reports are more appropriate to represent meaningful between-person differences (Finnigan & Vazire, 2018; Forbes et al., 2012; Oishi & Sullivan, 2006). For example, in the study by Oishi and Sullivan (2006), daily relationship satisfaction predicted later relationship status better than retrospective evaluations; however, the effect of daily relationship satisfaction was not incremental to global evaluations of relationship satisfaction. Studies applying a more continuous assessment or the Day Reconstruction Method (DRM, Kahneman, Krueger, Schkade, Schwarz, & Stone, 2014) might further help to disentangle which variance in retrospection can and which cannot be explained by actual experiences (but see Lucas, Wallsworth, Anusic, & Donnellan, 2019 for a critical comparison of ESM and DRM), as well as more studies examining the predictive power of each measure for different outcomes. In a first step in the current paper, however, the goal is to illustrate the degree of convergence between the different assessment modalities of relationship satisfaction. This requires to set one of both measures as reference category; in our case, we decided on the ESM state measures, but the research question could equally be examined using retrospection as reference category.

In the domain of intimate relationships, Fletcher and Kerr (2010) conducted a meta-analysis on the mean-level bias and accuracy of individuals' judgments. They differentiated six judgment categories, of which one dealt with retrospective evaluations of one's own assessments ("memories"). The authors report a positive mean-level bias for this category (i.e., an overestimation of relationship quality during retrospection); however, a closer look at the four studies that were included revealed that these studies dealt with different phenomena pertaining to a different interpretation of the mean-level bias. Specifically, three studies (Karney & Coombs, 2000; Karney & Frye, 2002; Sprecher, 1999) reported a positive mean-level bias of individuals' perception of *change* in relationship quality after time periods of 6 months to 10 years. A biased perception of change may differ from a biased perception of actual past experiences, because – depending on the concurrent assessment – a positively biased perception of change could mean a negatively biased perception of the actual experiences in the past. Indeed, a comparison of the level of relationship quality in retrospection with the *actual* assessment in the past indicates a negative mean-level bias in the studies of Karney and Coombs (2000) and Karney and Frye (2002; see also Holmberg and Holmes, 1994; Sprecher, 1999 did not examine retrospection of actual levels).

The fourth study that was included in the meta-analysis (Oishi & Sullivan, 2006) differed in some aspects from the other studies. First, the authors found a positive mean-level bias in retrospection with regard to actual past aspects of the relationship (i.e., not with regard to changes). Specifically, individuals overestimated the occurrence of partner-related behaviors (positive and negative ones), as well as their satisfaction for specific relationship domains in retrospection. Second, the retrospection occurred directly after a period of 14 days in which individuals rated these aspects of their relationship on a momentary basis. This difference in time between retrospection and experience across the studies included in the meta-analysis might be relevant for the bias that is occurring (see Robinson & Clore, 2002b; Walentynowicz, Schneider, & Stone, 2018 for effects of short vs. long time periods).

To summarize, the meta-analytic estimate of an overall positive mean-level bias for memories (Fletcher & Kerr, 2010) is a heterogeneous mix of findings which should not be interpreted without further consideration. In Study 1, we explored the mean-level bias of retrospective relationship satisfaction without any hypothesis in mind. Based on preliminary analyses in Study 1, for Study 2 we preregistered that we expect a negative mean-level bias (i.e., an underestimation of relationship satisfaction).

With regard to tracking accuracy, the meta-analysis of Fletcher and Kerr (2010) showed robust, significant and positive effects across all judgment categories. In line with these findings, we preregistered in both studies that we expect a positive association between the average ESM state and retrospection, translating into a positive tracking accuracy.

### What Moderates Mean-Level Bias? (RQ3)

A third goal of the current study concerned the exploration of possible moderators of a general mean-level bias. Regarding the retrospection of affective experiences, various moderators were identified in previous research, like personality (Feldman Barrett, 1997; Lay, Gerstorf, Scott, Pauly, & Hoppmann, 2017; Mill, Realo, & Allik, 2016), coping style (Schimmack & Hartmann, 1997), subjective well-being (Diener, Larsen, & Emmons, 1984), gender (Robinson, Johnson, & Shields, 1998), self-esteem (Christensen, Wood, & Feldman Barrett, 2003) or daily tiredness and age (Mill et al., 2016; Neubauer et al., 2019). The accessibility model of Robinson and Clore (2002a) suggests different sources of information individuals use when they report on their emotions. Momentary reports of individuals' emotions are described to be mainly driven by the experiential knowledge in the emotional situation, whereas retrospective reports shift from relying on accessible, episodic memory in short-term retrospection to relying on semantic memory and thereby to stable situation-specific or identity-related beliefs and heuristics in long-term retrospection (see Conner & Feldman Barrett, 2012 for a related account). This would explain why individual characteristics were found to moderate mean-level bias, when these are associated with beliefs about one's experiences and behavior in general (e.g., enhanced levels of remembered negative affect for individuals high in neuroticism, see Feldman Barrett, 1997; Lay et al., 2017; Mill et al., 2016).

Early research examining moderators of bias in the retrospection of relationship feelings indicates that individuals with low trust in their partner underestimate their own feelings for their partner (Holmberg & Holmes, 1994; see Luchies et al., 2013 for the role of trust in biased memories of the partner). The meta-analysis by Fletcher and Kerr (2010) also looked at moderators of mean-level biases and tracking accuracy. Bearing in mind that this meta-analysis was concerned with other judgment categories than memories as well, their results suggest that relationship quality, relationship length, and gender are important moderators for the mean-level bias observed across these different judgment categories. Specifically, individuals who are globally satisfied with their relationship seem to overall show an especially positive mean-level bias, although this relationship decreases with increasing length of the relationship. Attachment styles are also considered as potential influences (see also Pietromonaco & Feldman Barrett, 1997), which is in line with recent research showing that individuals overestimate their partner's negative emotions when they are high in attachment avoidance (Overall, Fletcher, Simpson, & Fillo, 2015).

Another line of research examined the influence of concurrent experiences on the biases occurring during retrospection. Two studies (Holmberg & Holmes, 1994; McFarland & Ross, 1987) found that relationship feelings during recall have an incremental effect on the retrospective assessment, in the way that the recall was similar to the present evaluation of the relationship (for a similar effect for mood and negative emotions see Chang, Overall, Madden, & Low, 2018; Parkinson et al., 1995).

In a longitudinal study covering three decades Karney and Coombs (2000) observed this pattern of consistency of retrospective assessments with current relationship satisfaction in a later stage of the relationship. These findings are in line with a theory by Ross (1989), which states that individuals reconstruct their autobiographical experiences based on their current status and then incorporating implicit theories of the malleability or stability of the experiences at hand. Such expectations may indeed play a role, as a study by Galak and Meyvis (2011) showed that individuals overestimate aversive experiences if they expect them to be repeated in the future.

In our studies, we thus explored individual differences that might invoke situation-specific or identity-related beliefs; global evaluations of the relationship or the partner; objective person and relationship characteristics; attachment styles; and concurrent global evaluations. As the current research focuses on the moderation of mean-level bias, we will shortly report, but not discuss the results concerning a moderation of tracking accuracy.

### What Level of Aggregation is Sufficient to Approach a Reliable Measurement of the Global Index? (RQ4)

Our last goal was to explore which number of ESM assessments of relationship satisfaction states account for what amount of variance of a global evaluation of relationship satisfaction. Epstein (1979) investigated a similar question for behavior, studying changes in reliability with an increasing number of daily behavioral assessments. The results showed that it takes around 14 days to achieve a satisfying correlation between behavioral samples of one person. For a time span up to four weeks, we will explore how strongly the association between the ESM assessments and the global index will rise with an increasing number of assessments, depending on the timing of the sampling (e.g., in the morning, evening, or a random survey during the day).

## Overview of Studies

For RQ2, the following hypotheses were preregistered: 1) For Study 1 (p. 8) and Study 2 (p. 41): "Individuals' relationship satisfaction retrospectively assessed after the experience sampling study is positively related to mean levels of individuals' state relationship satisfaction during the study (mean of states)." This translates to a positive tracking accuracy. 2) Only for Study 2 (p. 41): "Individuals' relationship satisfaction retrospectively assessed after the experience sampling study is lower than mean levels of individuals' state relationship satisfaction." This translates to a negative mean-level bias when regressing the retrospection on the average ESM states. We did not preregister how we were planning to analyze these specific hypotheses, but we preregistered some general exclusion criteria (see Sample), and how to handle multiple operationalizations (see Measures and **Table 1**). These preregistered decisions and deviations from them are highlighted accordingly in the respective sections. We did not have hypotheses concerning the performance of the different summary statistics in RQ1,[1] nor for RQ3 and RQ4, these analyses were exploratory.

Couples were recruited (via social networks, newsletters, flyers, notices at a German university and in Study 2 additionally with a website, and the help of therapists offering couple counseling) separately for two ESM studies with different study periods (14 days in Study 1, 28 days in Study 2). Requirements for participation were the affirmation to be at least 18 years old, to be in a heterosexual relationship with the declared partner, and to individually own an Android or iOS smartphone, which one could use regularly during the day. Participants provided a global evaluation of their relationship satisfaction and a range of other trait measures before they repeatedly rated their state relationship satisfaction five times a day. The studies finished with a retrospective assessment of the study period (and in Study 2 again with a more global evaluation of relationship satisfaction).

All measures were administered in German, if own translations were used, this is indicated accordingly. If not mentioned otherwise, for computation of scales, item responses were averaged. We used R (version 3.5.3, R Core Team, 2018) with the package *dplyr* for data handling (Wickham, François, Henry, & Müller, 2018), and the package *papaja* for manuscript writing (Aust & Barth, 2018). Both studies were part of a project funded by the German Research Foundation, which was approved by the local ethics committee. The data of Study 1 has previously

**Table 1:** Slider Items Used for the Assessment of State Relationship Satisfaction.

| Label | Question | Anchors | ESM | Retro |
|---|---|---|---|---|
| Item 1: Relationship Mood | How do you feel about your relationship at the moment? | bad (=0) over neutral (=5) to exceptionally good (=10) | Both Studies | Both Studies |
| Item 2: Annoyance (reverse) | How annoyed are you by your partner at the moment? | not at all (=0) to strongly (=10) | Both Studies | Only Study 2 |
| Item 3: Need Satisfaction | How are you feeling at the moment in your relationship? | frustrated (=0) over neutral (=5) to satisfied (=10) | Only Study 2 | Only Study 2 |
| Scale | Average of items | | Only Study 2 | Only Study 2 |

*Note*: Experience sampling items used for assessing state relationship satisfaction. The annoyance item was reverse coded for scale calculation. Please note that using only the relationship mood item for the analyses in Study 1 follows our preregistration. For Study 2, we preregistered to a) use the scale of all ESM relationship satisfaction items, but b) to use only relationship mood when it comes to retrospection; following these decisions would not allow for a commensurable comparison between the ESM measures and retrospection. Therefore, for Study 2, we report the results for all items and the scale separately (see main text for a more detailed description).

been used by Zygar et al. (2018a), the data of both studies by Pusch, Schönbrodt, Zygar-Hoffmann, and Hagemeyer (2019), as well as Schönbrodt, Zygar-Hoffmann, Nestler, Pusch, and Hagemeyer (2019). The results of these papers overlap with the analyses reported in the current paper only in basic descriptive statistics.[2]

## Study 1: Methods

### *Detailed Procedure*

Couples who signed up for the study could chose a time span of 13.5 hours (starting from 08:00 to 10:30 am, ending from 9:30 pm to midnight[3]) in which the daily, five ESM surveys were scheduled in a semi-random manner (approximately evenly distributed throughout the day) for a study period of 2 weeks. Next, individuals were invited to answer an online pre-ESM questionnaire on their personal computers (programmed with *formr*, Arslan & Tata, 2016; Arslan, Walther, & Tata, 2018) and received instructions for installing an ESM application on their own smartphones (developed at LMU Munich for Android devices). A personal login-code was assigned to each partner for matching the different data sets and identifying couples.

Right after logging into the ESM application, the questions and survey modalities were explained by written instructions, and the study period with in total 70 ESM surveys started on the day after the login. When a survey became active, individuals were notified by their smartphones and had 45 minutes to answer before the survey timed out. The median time needed to answer the survey was 3.28 minutes (interquartile range = 2.50). The questions were identical in each survey. Both partners were notified at the same time, but were asked to respond to the survey individually without discussing their answers with their partner.

After the ESM period, participants received a link to a post-ESM questionnaire (programmed with *LimeSurvey*, LimesurveyGmbH, 2017) which was to be answered on their personal computers. In this questionnaire individuals could also indicate if they wished to get a report on their answers and receive course credit. When their compliance was at least 80%, participants were also eligible to enter a raffle for a voucher. Due to a technical error, we could not retrieve the exact time difference between the end of the ESM part and the completion of the post-ESM questionnaire, but most participants completed the questionnaire within one to two weeks.

### *Sample*

The sample size in Study 1 was determined by time constraints: As we started data collection in November, we decided to finish it by the Christmas holidays to avoid potential bias during these special days. As one couple started two days later than planned and finished their study during the holidays, we excluded their answers on these days. Two persons participated although they were not in a relationship, so their entire data was excluded. This resulted in data from 152 individuals belonging to 77 couples for the pre-ESM questionnaire (two individuals participated without their partner).

We obtained data from a subset of 130 individuals from 68 couples for the ESM part of the study, as six couples quit after the pre-ESM questionnaire and two couples as well as six individuals answered less than the preregistered threshold of one third of all ESM surveys to be included in the final ESM sample (see p. 18 in the preregistration). Compliance for the everyday surveys was on average 84% (*SD* = 14%). After exclusion of 53 surveys for which participants reported that they had talked about their answers with their partner, the total number of (partly) answered measurement points was 7573.[4]

After the ESM study period, 117 individuals completed and one individual started (but did not finish) the post-ESM questionnaire. This sample consists of 66 women (56%), mainly students (83%), not married (97%) and without children (99%). For age and relationship duration, see **Table 2**, and for more details, see Zygar et al. (2018a).

### *Measures of Relationship Satisfaction*

#### Global relationship satisfaction (pre-ESM questionnaire)

For a global, holistic view on individuals' relationship satisfaction, we used the Couples Satisfaction Index (CSI (16); Funk & Rogge, 2007; Greischel & Johnson, n.d.) and the Positive-Negative Relationship Quality Scale (PNRQ, own translation; Rogge, Fincham, Crasta, & Maniaci, 2016). Whereas the CSI assesses global relationship satisfaction as an unidimensional construct, the PNRQ conceptualizes the evaluation of positive and negative qualities of the relationship as two separate constructs. In both measures, individuals are asked to rate their relationship regarding adjectives, but the CSI uses bipolar Likert scales (e.g., from 0 = *Boring* to 6 = *Interesting*), whereas the PNRQ presents single adjectives (e.g., "pleasant") which are to be evaluated on Likert scales ranging from 1 = *Not at all* to 7 = *Extremely*. The CSI additionally consists of questions such as "In general, how often do you think that things between you and your partner are going well?" with answers on 6- and 7-point Likert scales (see codebook for details). CSI ratings are summed.

#### State relationship satisfaction (ESM)

State relationship satisfaction was assessed with two questions (which we labeled "relationship mood" and "annoyance", see **Table 1**), with answers given on a continuous slider (without any slider ticks, without any numbers shown, results saved with multiple places after the decimal point, scale from 1 to 7 transformed to a 0–10 scale to match the scale of Study 2; see Schönbrodt et al., 2019 for an analysis of psychometric properties of these items). We considered these items to both reflect state relationship satisfaction, but as a minimum criterion for internal consistency on the between-moments level (also called event-level), we preregistered to only compute a scale if the event-level reliability exceeded .40 (see p. 17 in the preregistration). As this was not the case and because the retrospective assessment was only based on the relationship mood item, for Study 1 we only report results for this item.

#### Retrospective relationship satisfaction (post-ESM questionnaire)

In the post-ESM questionnaire individuals evaluated the two weeks of the ESM study period on the question "How did you overall feel about your relationship during these

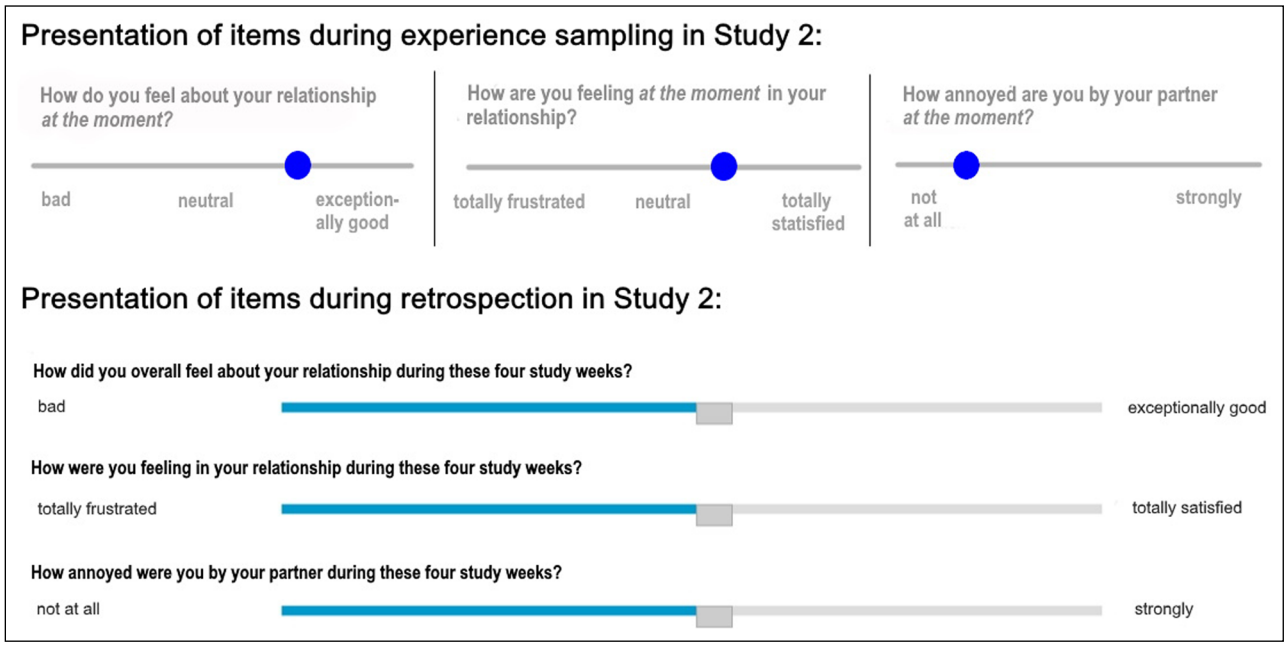


**Figure 1:** Presentation of items during experience sampling and during retrospection in Study 2 (translated). Presentation differed only slightly in Study 1. Figure available at https://osf.io/sq7mw/, under a CC-BY4.0 license.

two weeks?" with answers on a continuously presented slider ranging from *bad* (=0) to *exceptionally good* (=100; saved as whole numbers, linearly transformed to a 0–10 scale). There were three small differences compared to the state assessment, due to technical limitations (see **Figure 1**): a) There was no "neutral" label, which was present in the state assessment in the middle of the scale for the relationship mood item, b) The slider started in the middle, whereas no value was preselected in the state assessment, c) Whole numbers were shown as the slider was moved, which was not the case in the state assessment.

### *Potential Moderator Variables*

#### Personality (pre-ESM questionnaire)

The Big Five of personality were assessed with the 10-item short version of the Big Five Inventory (Rammstedt & John, 2007). Statements such as "I see myself as someone who gets nervous easily" (Neuroticism) were answered on a Likert scale (1 = *Disagree strongly*, 5 = *Agree strongly*).

#### Life satisfaction (pre-ESM questionnaire)

Individuals' overall satisfaction with their life was assessed with the Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985; Glaesmer, Grande, Braehler, & Roth, 2011). Participants rated five statements like "In most ways my life is close to my ideal." on a Likert scale (1 = *Strongly disagree*, 7 = *Strongly agree*).

#### Explicit social desires (pre-ESM questionnaire)

Explicit desires for affiliation, being alone and closeness were assessed with the ABC scale of social desires (Hagemeyer, Neyer, Neberich, & Asendorpf, 2013). Participants rated the frequency of 24 experiences related to social desires (e.g., "I enjoy it when my partner wants to be close to me." for closeness) on Likert scales (1 = *Never*, 7 = *Always*).

#### Intimacy in the relationship (pre-ESM questionnaire)

The amount of intimacy the participants experience in their relationship was measured with two self-constructed items. Individuals rated the frequency of events on questions such as "How often do you tell your partner what you are doing?" on a Likert scale (1 = *Never*, 5 = *Always*).

#### Further potential moderators (pre- and post-ESM questionnaire)

As moderators, we also examined person and relationship characteristics (gender, age, and relationship duration), dominance and autonomy in the relationship, self-reflection and insight (Grant, Franklin, & Langford, 2002), perception of the partner's explicit social desires (Hagemeyer et al., 2013), explicit motives (UMS-6; Schönbrodt & Gerstenberg, 2012), implicit partner-related needs (PACT; Hagemeyer & Neyer, 2012), and decision-making in the relationship (adaptation of the Allocation of Power in Decision-Making Areas Scale, Bell, 2008; Blood & Wolfe, 1960). As we did not find any effects for these variables as moderators, we refer to the Supplemental Materials and codebook for details.

## Study 2: Methods

### *Detailed Procedure*

For Study 2, the general study design was the same as in Study 1, with some exception in details: The ESM period lasted four instead of two weeks (with a total of 140 surveys), and couples were more flexible in their choice of the time span in which the surveys were scheduled. They could choose between a time span of 10 to 16 hours (starting from 07:00 to 10:00 am, ending from 9:00 pm to 11:00 pm) and could block up to two hours per day. A different ESM App was used, namely "Tellmi", which was developed at LMU Munich not only for Android but also for iOS devices. The questions and survey modalities were explained in a video upon login (instead of text-based in Study 1), and the study period started on the next Monday after the login

(instead of on the next day in Study 1). This time, the pre- *and* the post-ESM questionnaire were programmed with *formr* (Arslan et al., 2018; Arslan & Tata, 2017).

The medium time needed to answer the survey was 2.70 minutes (interquartile range = 2.17). The questions were identical for the first four surveys of the day. The evening survey differed with regard to the questions, and had a timeout of five hours instead of 45 minutes, because individuals were instructed to finish it before going to bed.

In addition to the opportunity of receiving a feedback report on their answers as in Study 1, participants were further compensated with course credit or money based on their compliance in the ESM part (up to 170€ per couple). In a follow-up questionnaire a year after the study couples could receive 20€ on top, and participate in a raffle for a voucher.

### *Sample*

Our sample size was constrained by the money available for participant compensation; 576 individuals belonging to 293 couples completed the pre-ESM questionnaire (10 individuals participated without their partner, these could not continue with the ESM part of the study).[5] We obtained data from a subset of 510 individuals from 259 couples for the ESM part, as six couples quit after the pre-ESM questionnaire and another 18 couples as well as eight individuals quit during the ESM part or answered less than the preregistered threshold of one third of all ESM surveys[6] to be included in the final ESM sample (after survey-level exclusions). Compliance for the everyday surveys of the remaining sample was on average 88% (*SD* = 12%). One couple changed time zone during the study but the survey timing did not adjust to the time transition, so in total 26 surveys (0.04%) were answered during the night and were excluded. As preregistered (see p. 59), we further excluded 171 surveys (0.24%) where individuals reported that they had talked about their answers with their partner and additional 1855 entries (2.58%) because of an answering time of less than 60 seconds. In total after all exclusions, 60942 (partly) answered measurement points remained.

After the ESM study period, 508 individuals completed the post-ESM questionnaire. However, we excluded the answers of 22 of these individuals for the retrospective assessment, because of apparently low quality data:[7] These individuals either did not change the default values that were preselected on all sliders ($n$ = 12) or probably overlooked the reverse coding of the annoyance item and were thus identified as outliers (Cook's Distance > 2 *SD*, $n$ = 10).[8] This resulted in a final sample of 486 individuals, consisting of 249 women (51%), mainly non-students (71%) without children (68%), with roughly one third of them married (32%); for age and relationship duration, see **Table 2**.

### *Measures of Relationship Satisfaction*

#### Global relationship satisfaction (pre-ESM questionnaire and post-ESM questionnaire)

We used the same measures as in Study 1 (CSI (16); Funk & Rogge, 2007, and PNRQ; Rogge et al., 2016), but also applied them in the post-ESM questionnaire, so we could examine the influence of concurrent relationship evaluations on the retrospective assessment.

#### State relationship satisfaction (ESM)

To achieve a more reliable assessment of state relationship satisfaction, we complemented the two items from Study 1 (but on a scale from 0–10) with an additional question with identical slider properties (which we called "need satisfaction", see **Table 1** and Schönbrodt et al., 2019). Again, as a minimum criterion for internal consistency, we preregistered to compute a scale if the event-level reliability exceeded .40, which was the case (see p. 42 in the preregistration).

#### Retrospective relationship satisfaction (post-ESM questionnaire)

In the post-ESM questionnaire individuals were asked to evaluate the study period on the questions presented in **Figure 1** with answers on a continuously presented slider with the same labels as for the state assessments (scale from 1 to 100, saved as whole numbers, again linearly transformed to a 0–10 scale). In contrast to Study 1, no numbers were shown as the slider was moved in the retrospective assessment, just as it was the case in the state assessment. Yet, two small differences compared to the state assessments remained (see **Figure 1**): As in Study 1, the "neutral" label was not shown in the retrospective assessment (which was present in the state assessment in the middle of the scale for the relationship mood and need satisfaction items), and the slider started in the middle of the scale instead of no default value being pre-selected.

Although for the retrospective assessment we had questions that were based on all three items, we preregistered to only use the relationship mood item (see p. 43 in the preregistration). To deal transparently with these inconsistencies in the preregistration regarding scale calculation of state and retrospective relationship satisfaction, for Study 2 we report the results for all three items and for the scale of all items separately, and correct accordingly for multiple comparisons. Next to providing transparency, this detailed presentation of the results a) allows to illustrate the cumulative evidence across both studies for the relationship mood item, which is the only item that was assessed both in Study 1 and Study 2 both in ESM and retrospection (see **Table 1**); b) informs which items are more susceptible to bias than others, therefore driving potential biases observed for the scale of all items.

### *Potential Moderator Variables*

We assessed the same moderator variables as in Study 1, but personality was assessed with another measure, attachment styles were additionally included and delay between the ESM period and retrospection was documented.

#### Personality (pre-ESM questionnaire)

The Big Five were measured with the 15 short-item scale developed for the Socio-Economic Panel survey (BFI-S; Gerlitz & Schupp, 2005). Participants rated statements such as "I see myself as someone who does a thorough

job" (Conscientiousness) on a Likert scale (1 = *Strongly disagree*, 7 = *Strongly agree*).

#### Attachment styles (pre-ESM questionnaire)

Anxiety and Avoidance in adult relationships were measured with the Experiences in Close Relationships Questionnaire (Ehrenthal, Dinger, Lamla, Funken, & Schauenburg, 2009). Thirty-six statements such as "I often worry that my partner doesn't really love me." (Anxiety) were answered on a Likert scale (1 = *Strongly disagree*, 7 = *Strongly agree*).

## Analysis Plan of Both Studies

In both studies state relationship satisfaction was measured repeatedly at the individual level, with individuals belonging to a specific dyad. To account for the resulting nonindependence of the data, we applied multilevel regression models (MLMs; using the packages *lme4* and *lmerTest*, Bates, Mächler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2017). In all models we entered a gender contrast as fixed effect (−1 = women, 1 = men, i.e., regression coefficients of other variables in the models can be interpreted as the average effect across both genders).[9] We aggregated the ESM data within individuals during preprocessing, hence individuals' summary of their ESM answers were on level 1 nested in couples on level 2. This pre-aggregation of ESM data was necessary to be able to compare summary statistics (for RQ1), and to be able to compute a slope while accounting for the nonindependence of the dyad data (for RQ2 and RQ3).

The relationship satisfaction variables (global/retrospective/aggregated state) were *z*-standardized for RQ1 to achieve a standardized regression coefficient, using the grand-mean and standard deviation across both genders. For the investigation of bias and accuracy (RQ2 and RQ3), the retrospective assessments and the aggregated ESM answers were grand-mean centered instead, using the grand-mean of the ESM measures (see West & Kenny, 2011): This results in both measures being centered on the variable that is conceptualized as the "truth" (i.e., the ESM answers). As both measures were transformed to the same metric, a mean-level bias would show itself in an intercept different from zero when regressing the retrospective assessment on the ESM answers. The sign of the intercept indicates whether the retrospective assessment is on average an under- or overestimation of the averaged feelings reported during ESM. The coefficient of the aggregated ESM measure shows the tracking accuracy, with a value of one representing perfect accuracy: An increase of one scale point in the aggregated ESM measure would then result in an increase of one scale point in the retrospective assessment. Entering moderators as main effects reveals whether individuals with a high expression of the moderator have an even higher or lower bias (i.e., conditional on the aggregated states as predictor, the main effect of the moderator variable increases or lowers the intercept). An interaction of the moderator variable with the aggregated ESM measure indicates whether tracking accuracy is decreased or increased for certain groups of individuals. The model including a moderator (i.e., for RQ3) is specified as follows (RQ2 uses the same model without all terms involving the moderator variable):

$$
\begin{aligned}
&Retrospection\left(GMC_{ESM}\right)_{ij} = \\
&\left(\gamma_{00} + \gamma_{10} Gender\ Contrast_{ij}\right. \\
&+ \gamma_{20} Moderator\left(z\right)_{ij} + \gamma_{30} Mean\ ESM\left(GMC_{ESM}\right)_{ij} \\
&\left. + \gamma_{40} Moderator\left(z\right)_{ij} \times Mean\ ESM\left(GMC_{ESM}\right)_{ij}\right) \\
&+ \left(u_{0j} + r_{ij}\right)
\end{aligned}
$$

with $GMC_{ESM}$ = grand-mean centered on the ESM-mean, $i$ = person-specific index, $j$ = couple-specific index, $\gamma$ = fixed effect, $(z)$ = $z$-standardized, $u$ = random intercept, $r$ = error term. This translates into the following between-person interpretation of the estimates:

$$
\begin{aligned}
&Retrospection = \\
&\left(Mean\text{-}Level\ Bias + Moderation\ of\ Bias\ by\ Gender\right. \\
&+ Moderation\ of\ Bias\ by\ Moderator + Tracking\ Accuracy \\
&\left. + Moderation\ of\ Accuracy\ by\ Moderator\right) \\
&+ \left(Random\ Intercept\ for\ Each\ Couple + Error\right)
\end{aligned}
$$

For all models, we report the marginal $R^2$ as an effect size, representing the explained variance by the fixed effects ($R^2_{GLMM(m)}$ from the *MuMIn* package, Johnson, 2014; Barton, 2018; Nakagawa & Schielzeth, 2013). When making multiple tests for a single analysis question (i.e., due to multiple items, summary statistics, moderators), we controlled the false discovery rate (FDR) at $\alpha$ = 5% (two-tailed) with the Benjamini-Hochberg (BH) correction of the *p*-values (Benjamini & Hochberg, 1995) implemented in the *stats* package (R Core Team, 2018).[10]

## Results of Both Studies

**Table 2** shows the descriptive statistics for both studies. Correlations and a complete description of the parameter estimates, confidence intervals, and effect sizes for all results can be found in the Supplemental Materials.

#### *What Summary Statistic Corresponds Best to Retrospection and Global Assessments? (RQ1)*

**Table 3** shows the standardized regression coefficients for several ESM summary statistics predicting retrospection after two weeks (Study 1) and four weeks (Study 2) of ESM, separately for the different relationship satisfaction items. For both studies and all items, the best prediction was achieved by the mean of the whole study period, while the mean of the last day and the 90th quantile of the distribution performed the worst. Overall, the highest associations were found for the mean of the scale of all three ESM items predicting the scale of all three retrospective assessments ($\beta$ = 0.75), and for the mean of need satisfaction predicting retrospection of this item ($\beta$ = 0.74).

The same analysis for the prediction of a global relationship satisfaction measure (the CSI) instead of the retrospective assessment is also shown in **Table 3** (for the prediction of PRQ and NRQ see Supplemental Materials).

**Table 2:** Descriptive Statistics.

| Variables | Study 1 | | | | Study 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha/\omega$ | *M* | *SD* | Range | $\alpha/\omega$ | *M* | *SD* | Range |
| Age in years | – | 22.44 | 4.29 | 18 to 40 | – | 31.29 | 9.49 | 18 to 68 |
| Relationship duration in years | – | 2.30 | 1.96 | 0 to 8 | – | 6.35 | 6.35 | 0.2 to 33.2 |
| Global RS: CSI (Pre-ESM) | 0.92 | 66.58 | 10.55 | 32 to 81 | 0.96 | 64.07 | 13.51 | 4 to 81 |
| Concurrent RS: CSI (Post-ESM) | – | – | – | – | 0.96 | 62.33 | 14.77 | 5 to 81 |
| Global RS: PRQ (Pre-ESM) | 0.92 | 5.83 | 0.98 | 1.5 to 7 | 0.91 | 5.74 | 0.88 | 2.4 to 7 |
| Concurrent RS: PRQ (Post-ESM) | – | – | – | – | 0.94 | 5.40 | 1.09 | 1 to 7 |
| Global RS: NRQ (Pre-ESM) | 0.91 | 1.86 | 1.05 | 1 to 5.9 | 0.94 | 1.84 | 1.07 | 1 to 7 |
| Concurrent RS: NRQ (Post-ESM) | – | – | – | – | 0.93 | 1.73 | 0.90 | 1 to 6.8 |
| Mean RS state: Item 1 | 0.95 | 7.03 | 1.14 | 3.3 to 9.8 | 0.98 | 7.25 | 1.34 | 2.7 to 10 |
| Retrospection of RS: Item 1 | – | 6.83 | 1.78 | 1.2 to 10 | – | 7.23 | 1.91 | 0 to 10 |
| Mean RS state: Item 2 (reverse) | 0.93 | 8.95 | 0.95 | 4.4 to 9.9 | 0.96 | 9.15 | 0.94 | 4.8 to 10 |
| Retrospection of RS: Item 2 (reverse) | – | – | – | – | – | 8.28 | 2.17 | 0 to 10 |
| Mean RS state: Item 3 | – | – | – | – | 0.98 | 7.20 | 1.38 | 1.5 to 10 |
| Retrospection of RS: Item 3 | – | – | – | – | – | 7.21 | 2.07 | 0 to 10 |
| Mean RS state: Scale | – | – | – | – | 0.97 | 7.86 | 1.11 | 3.1 to 10 |
| Retrospection of RS: Scale | – | – | – | – | 0.85 | 7.57 | 1.79 | 0.4 to 10 |
| Personality: Conscientiousness | 0.49 | 3.50 | 0.83 | 1.5 to 5 | 0.69 | 5.25 | 1.09 | 1 to 7 |
| Personality: Neuroticism | 0.62 | 2.89 | 1.10 | 1 to 5 | 0.68 | 4.14 | 1.34 | 1 to 7 |
| Satisfaction with life | 0.88 | 5.50 | 1.07 | 2 to 7 | 0.87 | 5.16 | 1.16 | 1.2 to 7 |
| Explicit desire for being alone | 0.84 | 4.07 | 0.94 | 1.8 to 6.4 | 0.85 | 4.21 | 0.98 | 1 to 6.9 |
| Explicit desire for closeness | 0.86 | 6.18 | 0.67 | 3.5 to 7 | 0.90 | 6.03 | 0.76 | 1.4 to 7 |
| Intimacy in the relationship | 0.79 | 4.12 | 0.78 | 1.5 to 5 | 0.82 | 3.78 | 0.89 | 1.5 to 5 |
| AS: Anxiety in the relationship | – | – | – | – | 0.90 | 2.81 | 1.08 | 1 to 6.2 |
| AS: Avoidance in the relationship | – | – | – | – | 0.89 | 2.22 | 0.85 | 1 to 6.4 |
| Delay of retrospection in days | – | – | – | – | – | 2.01 | 4.05 | 0 to 63 |

*Note*: N (Study 1) = 118–152, N (Study 2) = 486–576, RS = Relationship Satisfaction, CSI = Couples Satisfaction Index, PRQ = Positive Relationship Quality, NRQ = Negative Relationship Quality, Item 1 = Relationship mood, Item 2 = Annoyance (reverse coded), Item 3 = Need satisfaction, AS = Attachment Style. For state measures the between-person reliability is reported, for scales consisting of only two items Cronbach's $\alpha$ is reported, and for all other measures McDonald's $\omega_{total}$ is reported.

The mean of the last week, of the last day and of the first week were not entered as predictors, as they provide no special meaning to the global evaluation, which was assessed before the ESM part. Again, the mean was the best predictor in all cases. Other summary statistics performed equally well in some cases, but without a systematic pattern. The associations were highest when the mean of the scale, or the mean of need satisfaction (item 3) across four weeks predicted the CSI ($\beta_{Scale} = 0.59$, $\beta_{NeedSatisfaction} = 0.58$).

We additionally checked whether other summary statistics next to the mean provided an incremental contribution to the prediction of retrospection (see **Table 4**). This was not the case in Study 1 (we controlled the FDR for all incremental effects across studies, all BH-corrected *p*s of the model comparisons >0.16). In Study 2, all summary statistics except the 90th quantile and the mean of the first week made incremental contributions for the prediction of retrospection of relationship mood and the scale. For the annoyance item both the 10th and the 90th quantile – but no other summary statistic – had incremental effects. As annoyance was reverse coded, the 10th quantile represents a high level of annoyance, whereas the 90th quantile represents a low level of annoyance. For need satisfaction only the summaries of the end of the study (i.e., mean of the last week and mean of the last day) had additional relevance. Overall the incremental contributions were small (additional explained variance <3%, compared to baseline explained variance of the mean as single predictor between 30% and 57%). Whereas the coefficients of the 10th quantile and the means of the last day/week were positive, the median and the 90th quantile had negative coefficients.

**Table 3:** Prediction of Retrospective and Global Assessment by Different Summary Statistics of ESM Relationship Satisfaction States (all *z*-Standardized).

| | Study 1 | Study 2 | | | |
|---|---|---|---|---|---|
| **Retrospection by summary** | **Item 1** | **Item 1** | **Item 2 (r)** | **Item 3** | **Scale** |
| Mean | 0.55 | 0.66 | 0.61 | 0.74 | **0.75** |
| Mean last week | 0.54 | 0.65 | 0.55 | 0.72 | 0.72 |
| 10th quantile | 0.53 | 0.63 | 0.59 | 0.68 | 0.71 |
| Median | 0.52 | 0.61 | 0.49 | 0.70 | 0.70 |
| Mean first week | 0.48 | 0.56 | 0.52 | 0.64 | 0.65 |
| Mean last day | 0.43 | 0.57 | 0.41 | 0.62 | 0.59 |
| 90th quantile | 0.40 | 0.54 | 0.28 | 0.61 | 0.60 |
| **CSI by summary** | **Item 1** | **Item 1** | **Item 2 (r)** | **Item 3** | **Scale** |
| Mean | 0.38 | 0.52 | 0.44 | 0.58 | **0.59** |
| 10th quantile | 0.38 | 0.50 | 0.41 | 0.52 | 0.55 |
| Median | 0.33 | 0.46 | 0.35 | 0.52 | 0.54 |
| 90th quantile | 0.30 | 0.50 | 0.25 | 0.55 | 0.55 |

*Note*: N (Study 1) = 115–130, N (Study 2) = 475–510. Item 1 = Relationship mood, Item 2 = Annoyance (reverse coded), Item 3 = Need satisfaction. CSI = Couples Satisfaction Index assessed before the ESM period. Rows ordered by size of average coefficient across all items. The strongest effect is printed in bold.

**Table 4:** Prediction of Retrospection by Relationship Satisfaction States: Incremental Contributions Beyond the Mean.

| | Study 1 | Study 2 | | | |
|---|---|---|---|---|---|
| | **Relationship mood** | **Relationship mood** | **Annoyance (reverse)** | **Need satisfaction** | **Scale** |
| Baseline $R^2$ | 30.21% | 43.54% | 36.95% | 56.65% | 56.65% |
| Mean | b = 0.63, p = .039 | b = 1.25, p < .001 | b = 0.77, p < .001 | b = 0.98, p < .001 | b = 1.10, p < .001 |
| Median | b = –0.08, p = .793 | **b = −0.60, p < .001** | b = –0.18, p = .037 | b = –0.24, p = .091 | **b = −0.35, p = .014** |
| Δ $R^2$ | 0.02% | **1.96%** | 0.70% | 0.46% | **0.86%** |
| Mean | b = 0.42, p = .008 | b = 0.42, p < .001 | b = 0.44, p < .001 | b = 0.63, p < .001 | b = 0.58, p < .001 |
| 10q | b = 0.15, p = .308 | **b = 0.27, p < .001** | **b = 0.20, p = .018** | b = 0.12, p = .059 | **b = 0.19, p = .006** |
| Δ $R^2$ | 1.13% | **1.93%** | **0.93%** | 0.52% | **1.01%** |
| Mean | b = 0.77, p < .001 | b = 0.67, p < .001 | b = 0.71, p < .001 | b = 0.79, p < .001 | b = 0.82, p < .001 |
| 90q | b = –0.26, p = .071 | b = –0.02, p = .762 | **b = −0.16, p = .001** | b = –0.05, p = .366 | b = –0.09, p = .140 |
| Δ $R^2$ | 2.71% | –0.05% | **1.48%** | 0.05% | 0.30% |
| Mean | b = 0.54, p < .001 | b = 0.54, p < .001 | b = 0.58, p < .001 | b = 0.63, p < .001 | b = 0.66, p < .001 |
| Mean last day[1] | b = 0.00, p = .970 | **b = 0.16, p = .003** | b = 0.07, p = .115 | **b = 0.16, p < .001** | **b = 0.13, p = .003** |
| Δ $R^2$ | 0.17% | **1.24%** | 0.14% | **1.22%** | **0.87%** |
| Mean | b = 0.32, p = .193 | b = 0.39, p < .001 | b = 0.53, p < .001 | b = 0.49, p < .001 | b = 0.51, p < .001 |
| Mean last week[1] | b = 0.24, p = .327 | **b = 0.29, p = .002** | b = 0.09, p = .241 | **b = 0.27, p = .001** | **b = 0.25, p = .002** |
| Δ $R^2$ | 0.61% | **1.24%** | 0.10% | **0.81%** | **0.86%** |
| Mean | b = 0.85, p < .001 | b = 0.73, p < .001 | b = 0.56, p < .001 | b = 0.86, p < .001 | b = 0.74, p < .001 |
| Mean first week | b = –0.32, p = .170 | b = –0.08, p = .321 | b = 0.06, p = .394 | b = –0.13, p = .083 | b = 0.00, p = .971 |
| Δ $R^2$ | 1.23% | 0.09% | 0.04% | 0.24% | –0.06% |

*Note*: N (Study 1) = 118, N (Study 2) = 486. Baseline $R^2$ is the explained variance by the mean as fixed effect. Δ $R^2$ is the incremental explained variance by the additional summary statistic, compared to the model including only the mean as predictor. [1]Due to missing data on the last day or last week for some persons, these models used data from only 115 participants in Study 1 (for models with the last day) and from 475/485 participants in Study 2 (for models with the last day/last week); the baseline $R^2$ differs slightly on this data. Bold values of the additional summary statistics indicate that a model without this variable fits the data significantly worse after controlling the false discovery rate at $\alpha = 5\%$ (two-tailed) for all model comparisons. The predictors and the criterion in the models are *z*-standardized. The 10th quantile represents an especially negative relationship evaluation for all items (as annoyance is reverse coded); the 90th quantile represents an especially positive relationship evaluation for all items. Please note that mean and median very highly correlated, leading to Variance Inflation Factors (VIFs) between 5 and 23; all other VIFs were <10.

### *What Bias Occurs in Retrospection? (RQ2)*

Given that the mean was the best measure for predicting retrospection, for investigating mean-level bias and tracking accuracy, we regressed the retrospective assessment on the mean of relationship satisfaction states. **Table 5** shows the results for the different items, including a meta-analytical *p*-value for the relationship mood item (calculated with the *metap* package, Dewey, 2018), to synthesize the results of both studies.

There was no significant mean-level bias for the two positively framed items (relationship mood and need satisfaction). However, for the negatively framed annoyance item and for the scale out of all three items, a negative mean-level bias emerged.[11] It is important to note that the annoyance item was reverse coded, therefore the negative coefficient of the mean-level bias indicates that individuals on average *overestimate* the amount of them having been annoyed by their partner during the study.[12] This bias is still present when computing the scale that includes annoyance next to relationship mood and need satisfaction. In consequence, individuals' overall relationship satisfaction score is lower in retrospection than the average ESM report, driven by a higher level of remembered annoyance.[13]

**Table 5:** Prediction of Retrospective Assessment by Mean of ESM Relationship Satisfaction States (With Common Zero).

| | Relationship mood (S1) | | | Relationship mood (S2) | | | Relationship mood (S1 + S2) |
|---|---|---|---|---|---|---|---|
| | ***β*** | **95% CI** | ***p*** | ***β*** | **95% CI** | ***p*** | **meta *p*** |
| Intercept (Mean-level Bias) | –0.19 | [–0.49,0.12] | .237 | –0.01 | [–0.15,0.13] | .855 | .736 |
| Gender | 0.04 | [–0.19,0.26] | .737 | –0.05 | [–0.16,0.07] | .411 | .532 |
| ESM Mean (Tracking Accuracy) | 0.86 | [0.62,1.10] | .247 | 0.93 | [0.83,1.03] | .175 | .127 |
| $R^2_{GLMM(m)}$ | | .302 | | | .435 | | |
| | **Annoyance (reverse) (S2)** | | | **Need satisfaction (S2)** | | | **Scale (S2)** |
| | ***β*** | **95% CI** | ***p*** | ***β*** | **95% CI** | ***p*** | ***β*** | **95% CI** | ***p*** |
| Intercept (Mean-level Bias) | –0.87 | [–1.03,–0.71] | <.001 | 0.01 | [–0.12,0.14] | .864 | –0.29 | [–0.40,–0.18] | <.001 |
| Gender | 0.13 | [–0.02,0.28] | .087 | –0.09 | [–0.19,0.02] | .114 | –0.01 | [–0.10,0.09] | .900 |
| ESM Mean (Tracking Accuracy) | 1.39 | [1.22,1.55] | <.001 | 1.12 | [1.03,1.21] | .011 | 1.20 | [1.10,1.29] | <.001 |
| $R^2_{GLMM(m)}$ | | .369 | | | .566 | | | .567 | |

*Note*: S1 = Study 1, S2 = Study 2, Gender = Contrast variable with –1 = women and 1 = men, CI = Confidence Interval. N (Study 1) = 118, N (Study 2) = 486. Retrospective assessment and mean of states were centered on the grand-mean of the mean of states. The intercept of the models indicate whether mean-level bias is present, the slope of the ESM mean state indicates whether the tracking accuracy differs from 1 (likewise, we tested whether the slope differs from 1, i.e., the *p*-value corresponds to the H0: $\beta = 1$). All significant *p*-values remain significant after controlling the false discovery rate at $\alpha = 5\%$ (two-tailed).

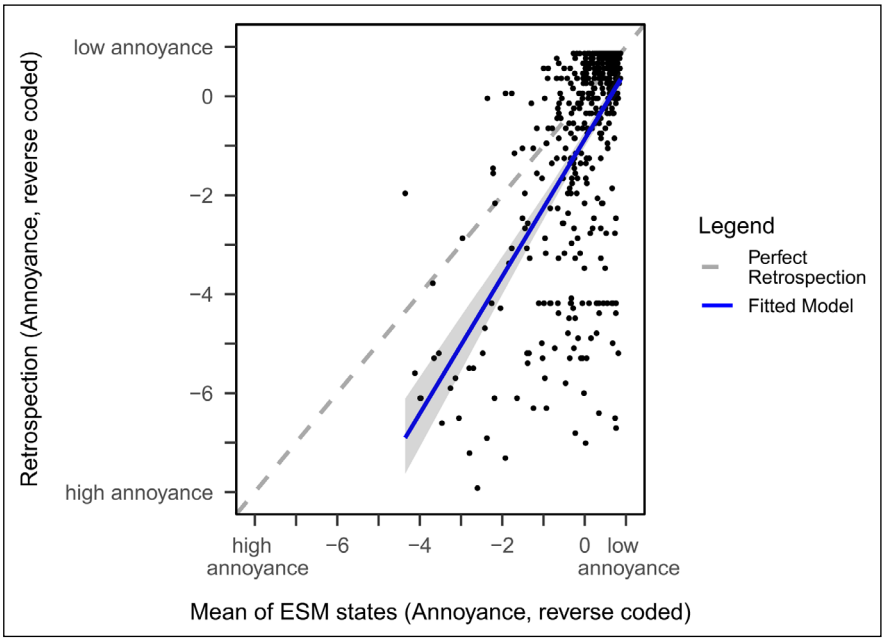


**Figure 2:** Prediction of retrospective assessment by mean of ESM relationship satisfaction states for the reverse coded annoyance item (with common zero). High values indicate low annoyance. Uncertainty band was calculated with the *merTools* package (Knowles & Frederick, 2018). Figure created with the *ggplot2* package (Wickham, 2016), available at https://osf.io/sq7mw/, under a CC-BY4.0 license.

Further, the results showed a tracking accuracy of greater than one for the annoyance and need satisfaction item and for the scale. This indicates that experienced annoyance captured by the ESM assessments is amplified during retrospection: High levels of being annoyed are perceived as having been even higher, reinforcing the negative mean-level bias, and leading to an overall more diverging perception. For low annoyance, this effect counterbalances the mean-level bias and results in an overall more similar perception (see **Figure 2**).[14]

### *What Moderates Mean-Level Bias? (RQ3)*

We added moderators of mean-level bias and tracking accuracy to the models of RQ2, so that retrospection was predicted by an intercept (indicating potential mean-level bias), a main effect of the mean ESM state (indicating potential tracking accuracy), a main effect of a moderator (indicating a potential moderation of the mean-level bias) and the interaction between mean ESM state and the moderator (indicating a potential moderation of the tracking accuracy). We report the results of those moderators that had a significant main effect for at least one item or the scale after controlling the FDR.

**Figure 3** illustrates the pattern of main effects for global relationship satisfaction as a moderator: Independent of the item being considered, global relationship satisfaction concurrently assessed with retrospection turned out to be a central moderator of the mean-level bias in both studies, irrelevant of the measure being the CSI or the more specific PNRQ scales. The coefficients indicate that individuals who are globally more satisfied with their relationship during retrospection tend to less strongly underestimate or even overestimate their relationship satisfaction as reported during ESM. In case of annoyance, due to the reverse coding, the coefficients indicate that globally satisfied individuals less strongly overestimate their level

**Results of models with different moderators, separately for different items**

| | Intercept | | Moderator | |
|---|---|---|---|---|
| Variables | beta | p | beta | p |
| **Moderator: Concurrent CSI** | | | | |
| Relationship mood (S2) | 0.03 | .690 | 0.66 | < .001 |
| Annoyance (reverse) (S2) | −0.95 | < .001 | 0.59 | < .001 |
| Need satisfaction (S2) | 0.02 | .805 | 0.62 | < .001 |
| Scale | −0.30 | < .001 | 0.61 | < .001 |
| **Moderator: Concurrent PRQ** | | | | |
| Relationship mood (S2) | 0.07 | .331 | 0.57 | < .001 |
| Annoyance (reverse) (S2) | −0.94 | < .001 | 0.41 | < .001 |
| Need satisfaction (S2) | 0.06 | .412 | 0.52 | < .001 |
| Scale | −0.26 | < .001 | 0.45 | < .001 |
| **Moderator: Concurrent NRQ** | | | | |
| Relationship mood (S2) | 0.05 | .500 | −0.47 | < .001 |
| Annoyance (reverse) (S2) | −0.98 | < .001 | −0.54 | < .001 |
| Need satisfaction (S2) | 0.03 | .668 | −0.50 | < .001 |
| Scale | −0.30 | < .001 | −0.50 | < .001 |
| **Moderator: Pre−ESM CSI** | | | | |
| Relationship mood (S1) | −0.15 | .359 | 0.32 | .049 |
| Relationship mood (S2) | 0.06 | .451 | 0.32 | < .001 |
| Annoyance (reverse) (S2) | −0.95 | < .001 | 0.42 | < .001 |
| Need satisfaction (S2) | 0.05 | .445 | 0.20 | .022 |
| Scale | −0.27 | < .001 | 0.27 | < .001 |
| **Moderator: Pre−ESM PRQ** | | | | |
| Relationship mood (S1) | −0.19 | .252 | −0.12 | .405 |
| Relationship mood (S2) | 0.04 | .565 | 0.26 | .001 |
| Annoyance (reverse) (S2) | −0.91 | < .001 | 0.25 | .004 |
| Need satisfaction (S2) | 0.05 | .506 | 0.12 | .106 |
| Scale | −0.27 | < .001 | 0.15 | .020 |
| **Moderator: Pre−ESM NRQ** | | | | |
| Relationship mood (S1) | −0.17 | .299 | 0.06 | .710 |
| Relationship mood (S2) | 0.03 | .629 | −0.14 | .077 |
| Annoyance (reverse) (S2) | −0.93 | < .001 | −0.16 | .071 |
| Need satisfaction (S2) | 0.04 | .533 | −0.08 | .312 |
| Scale | −0.29 | < .001 | −0.11 | .101 |

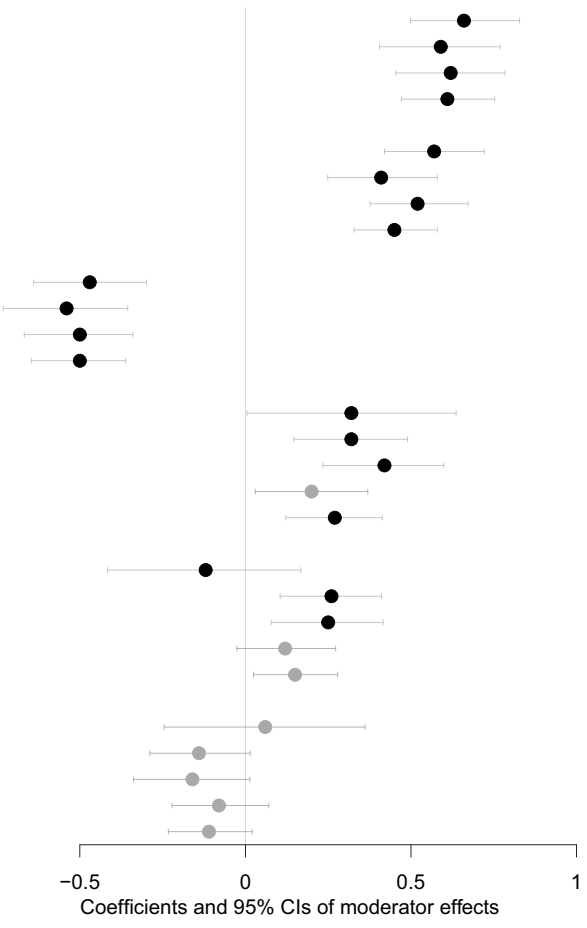


**Figure 3:** Moderation of mean-level bias by global relationship satisfaction (i.e., main effects of global relationship satisfaction concurrently assessed and assessed "pre-esm" = before the experience sampling study) for different relationship satisfaction items. The interaction between moderator and mean relationship satisfaction states (i.e., the moderation of tracking accuracy) is included in the models, but not reported here. S1 = Study 1, S2 = Study 2. N (Study 1) = 118, N (Study 2) = 486. Moderator effects that were significant after controlling the false discovery rate at $\alpha = 5\%$ (two-tailed) are displayed in black (for relationship mood based on a meta *p*-value of both studies), all other moderator effects are displayed in grey. Figure created with the *forestplot* package (Gordon & Lumley, 2017), available at https://osf.io/sq7mw/, under a CC-BY4.0 license.

of annoyance. Even though the overall mean-level bias for the relationship mood and need satisfaction items was not significantly different from zero (see RQ2 and "Intercept" column in **Figure 3**), the models with these items still showed the moderating effect by the global measure.

Global relationship satisfaction assessed before the evaluated ESM period had similar, but considerably lower and more inconsistent effects: The aforementioned moderation was present for all items except need satisfaction when looking at the CSI; the moderation by the PRQ was only significant for the annoyance and the need satisfaction item; and there was no significant moderation by the NRQ.

As shown in **Figure 4**, life satisfaction had likewise a positive moderating effect for all items, indicating that individuals who are globally happy with their life show

**Results of models with different moderators, separately for different items**

| | Intercept | | Moderator | |
|---|---|---|---|---|
| Variables | beta | p | beta | p |
| **Moderator: Satisfaction with life** | | | | |
| Relationship mood (S1) | −0.20 | .195 | 0.19 | .150 |
| Relationship mood (S2) | 0.00 | .973 | 0.17 | .009 |
| Annoyance (reverse) (S2) | −0.88 | < .001 | 0.27 | .001 |
| Need satisfaction (S2) | 0.03 | .701 | 0.19 | .002 |
| Scale | −0.27 | < .001 | 0.18 | .001 |
| **Moderator: Anxious AS** | | | | |
| Relationship mood (S2) | 0.00 | .954 | −0.12 | .067 |
| Annoyance (reverse) (S2) | −0.87 | < .001 | −0.27 | .001 |
| Need satisfaction (S2) | 0.02 | .762 | −0.12 | .068 |
| Scale | −0.28 | < .001 | −0.15 | .008 |
| **Moderator: Avoidant AS** | | | | |
| Relationship mood (S2) | 0.04 | .549 | −0.23 | .001 |
| Annoyance (reverse) (S2) | −0.87 | < .001 | −0.26 | .003 |
| Need satisfaction (S2) | 0.05 | .437 | −0.09 | .207 |
| Scale | −0.24 | < .001 | −0.14 | .018 |
| **Moderator: Desire for being alone** | | | | |
| Relationship mood (S1) | −0.16 | .273 | −0.30 | .032 |
| Relationship mood (S2) | 0.00 | .948 | −0.15 | .020 |
| Annoyance (reverse) (S2) | −0.87 | < .001 | 0.10 | .208 |
| Need satisfaction (S2) | 0.01 | .904 | 0.00 | .989 |
| Scale | −0.29 | < .001 | −0.02 | .692 |
| **Moderator: Neuroticism** | | | | |
| Relationship mood (S1) | −0.18 | .250 | −0.21 | .103 |
| Relationship mood (S2) | −0.01 | .861 | −0.03 | .668 |
| Annoyance (reverse) (S2) | −0.86 | < .001 | −0.21 | .011 |
| Need satisfaction (S2) | 0.01 | .849 | −0.14 | .036 |
| Scale | −0.29 | < .001 | −0.11 | .047 |
| **Moderator: Desire for closeness** | | | | |
| Relationship mood (S1) | −0.24 | .150 | 0.21 | .187 |
| Relationship mood (S2) | 0.00 | .986 | 0.15 | .059 |
| Annoyance (reverse) (S2) | −0.92 | < .001 | 0.26 | .002 |
| Need satisfaction (S2) | 0.01 | .899 | 0.02 | .784 |
| Scale | −0.29 | < .001 | 0.10 | .107 |
| **Moderator: Intimacy** | | | | |
| Relationship mood (S1) | −0.13 | .424 | 0.03 | .853 |
| Relationship mood (S2) | 0.02 | .760 | 0.09 | .191 |
| Annoyance (reverse) (S2) | −0.93 | < .001 | 0.25 | .003 |
| Need satisfaction (S2) | 0.06 | .411 | 0.09 | .194 |
| Scale | −0.28 | < .001 | 0.12 | .045 |
| **Moderator: Conscientiousness** | | | | |
| Relationship mood (S1) | −0.20 | .222 | −0.09 | .481 |
| Relationship mood (S2) | −0.01 | .876 | 0.02 | .716 |
| Annoyance (reverse) (S2) | −0.86 | < .001 | 0.21 | .008 |
| Need satisfaction (S2) | 0.02 | .805 | −0.02 | .739 |
| Scale | −0.28 | < .001 | 0.06 | .287 |
| **Moderator: General (pos/neg) factor** | | | | |
| Relationship mood (S2) | 0.03 | .642 | 0.28 | < .001 |
| Annoyance (reverse) (S2) | −0.90 | < .001 | 0.32 | .001 |
| Need satisfaction (S2) | 0.04 | .546 | 0.12 | .113 |
| Scale | −0.26 | < .001 | 0.19 | .005 |

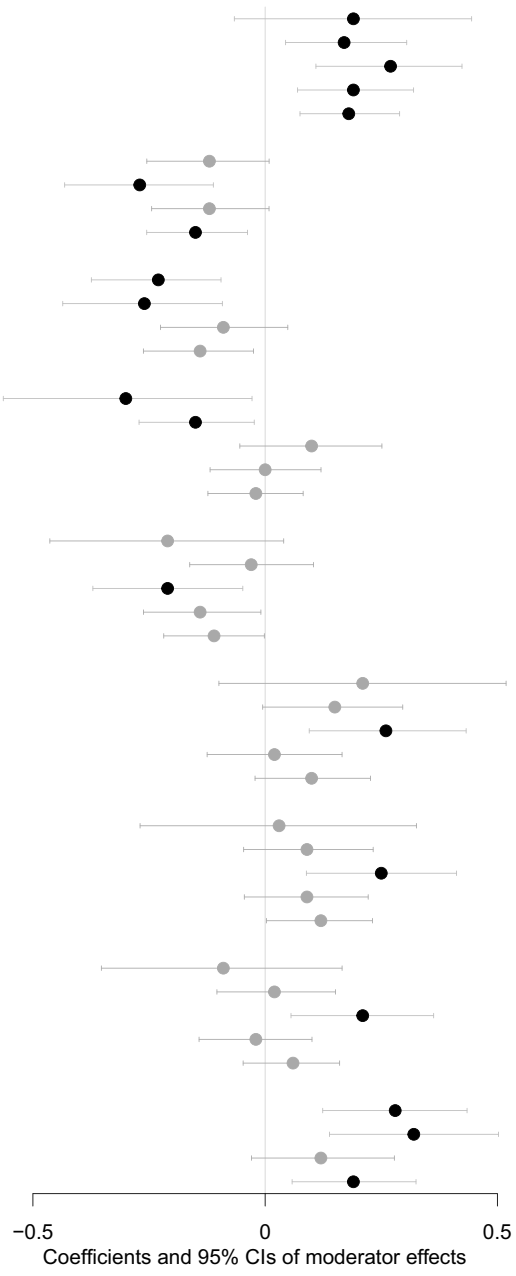


**Figure 4:** Moderation of mean-level bias by different moderators (i.e., main effects of these moderators) for different relationship satisfaction items. The interaction between moderator and mean relationship satisfaction states (i.e., the moderation of tracking accuracy) is included in the models but not reported here. S1 = Study 1, S2 = Study 2. N (Study 1) = 118, N (Study 2) = 486, AS = Attachment Style. Moderator effects that were significant after controlling the false discovery rate at $\alpha = 5\%$ (two-tailed) are displayed in black (for relationship mood based on a meta *p*-value of both studies), all other moderator effects are displayed in grey. Figure created with the *forestplot* package (Gordon & Lumley, 2017), available at https://osf.io/sq7mw/, under a CC-BY4.0 license.

less of an overall underestimation of their relationship satisfaction, resulting from a less strongly overestimation of annoyance and some overestimation of relationship mood and need satisfaction. In contrast, anxious and avoidant attachment, neuroticism, and the explicit desire for being alone had negative moderating effects on some items. Individuals with a high expression of these traits underestimate their relationship satisfaction in some aspects even stronger.

There were some other moderators that only influenced the bias of specifically the annoyance item: The explicit desire for closeness, perceived intimacy, and conscientiousness all had positive effects, counterbalancing the overall negative bias in the evaluation of annoyance (i.e., resulting in a less strongly overestimation for those scoring high on these traits; see **Figure 4**).[15]

The result pattern suggests that all moderators with positive valence show a positive moderating effect, and those with negative valence a negative effect. Consequently, these findings could result from an overall latent factor reflecting positive compared to negative views about oneself/one's life/one's relationship or more generally a methodological artefact of social desirability. As a first approach to this alternative explanation, we fitted a bifactor model (see e.g., Biderman, Nguyen, Cunningham, & Ghorbani, 2011; Reise, 2012) with structural equation modeling (using *lavaan*, Rosseel, 2012) on all self-report items assessed during the pre-ESM questionnaire in Study 2: In this model all items load on their respective scales (with correlated latent factors of all these scales), as well as on a general factor (orthogonal to the other latent factors). The general factor that resulted from this analysis seems to capture indeed a general positivity or negativity in answering the items (i.e., all items from constructs mirroring positive feelings or experiences loaded positively, irrespective of them being reverse scored or not; items from constructs reflecting negative feelings or experiences loaded negatively; model fit and all factor loadings are presented in the Supplemental Materials). In a second step, we extracted regression factor scores on this latent factor for each person, and added them as additional manifest moderator variable to our analyses (see **Figure 4**): The results show that this factor moderates the mean-level bias of relationship mood, annoyance, and the scale, but not of need satisfaction.

To assess whether the specific moderators explain variance beyond this general positivity factor, we repeated all analyses with this factor included as covariate (as main effect and in interaction with the averaged ESM states). Robust to adding this control variable were the moderation effects of all relationship satisfaction measures concurrently assessed; of the CSI assessed before the ESM study period; of life satisfaction on all but the relationship mood item; of anxious attachment and conscientiousness on the annoyance item (uncorrected *p*-values of these moderators <.05). Not robust were the effects of the PRQ measured before the study period; of life satisfaction on the relationship mood item; of anxious attachment on the scale; of avoidant attachment and neuroticism; and of intimacy, the explicit desires for closeness and for being alone on the annoyance item.

The tracking accuracy was moderated only by the intimacy in the relationship and concurrent negative relationship quality for some items (see Supplemental Materials).

### *What Level of Aggregation is Sufficient to Approach a Reliable Measurement of the Global Index? (RQ4)*

For RQ4 we only report the results for Study 2 in the main text, because in this study four instead of only two weeks of sampling were available. The respective results for Study 1 can be found in the Supplemental Materials. **Figure 5** shows the association between different numbers and schedules of ESM assessments and the CSI as global relationship satisfaction measure assessed before the ESM study period. Using all five assessments of the day for all four weeks that were sampled, the association between the aggregated ESM state relationship satisfaction scale

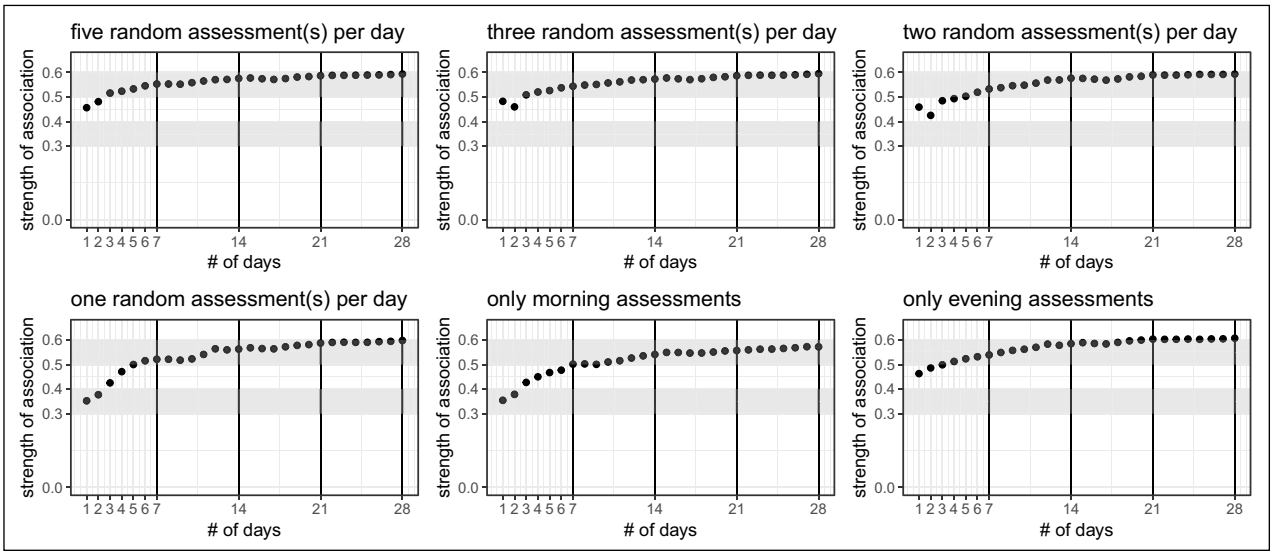


**Figure 5:** Association between (aggregated) state relationship satisfaction and global relationship satisfaction for different number of assessments and schedules in Study 2.

and the CSI was $\beta$ = .59 (see **Table 3**). The size of the association was already nearly achieved after one ($\beta$ = .55) or two weeks of sampling ($\beta$ = .57).

Looking at different numbers of assessments per day with a random sampling plan shows in both studies that a higher number of assessments matters only for the first few days. Afterwards, a higher sampling rate does not increase the effect size of the association meaningfully faster or stronger than fewer assessments.

Comparing evening assessments with morning and single random assessments shows in Study 2 that the evening assessments descriptively reach peak associations slightly sooner than the other sampling plans. However, we could not observe similar differences between the sampling plans in Study 1.

## Discussion

The present studies tapped into different aspects of assessing relationship satisfaction, comparing state assessments with retrospective assessments and global evaluations. To understand the relationship between states, global and retrospective evaluations, different summary statistics of the state assessments were evaluated in their ability to predict the other assessment modes. Averaging the state assessments showed the highest association with the other two measures in both studies, but most other summary statistics performed similarly well or provided small incremental information. When individuals try to recap their experiences in their relationship, they might remember some occurrences better than other ones. We therefore compared the retrospective assessments with the averaged state reports to assess tracking accuracy and to uncover a potential mean-level bias of the sample when recalling the study weeks. As expected, the resulting tracking accuracy was positive, confirming that individuals' retrospective assessments converge to a large extent with what they on average report to have experienced on a momentary basis; however, the estimation differed significantly from a perfect tracking accuracy of one for all but the relationship mood item, indicating also the presence of systematic deviations. We further found a negative mean-level bias during retrospection for the scale of all items in Study 2, driven by individuals reporting a stronger intensity of them having been annoyed in their relationship compared to the average of what they indiciated on a momentary basis.

We explored several moderators of this mean-level bias, and found the strongest to be global relationship satisfaction concurrently assessed with the retrospection: Individuals who are globally more satisfied with their relationship when they recall their study weeks, tend to less strongly overestimate their level of annoyance, and also tend to indicate retrospectively better relationship mood and need satisfaction in the relationship. This moderating effect was also observed for global relationship satisfaction assessed before the study period, albeit less strongly and not for all measures, as well as for individuals who report higher levels of life satisfaction, intimacy in their relationship, desire for closeness, and conscientiousness. Individuals who showed higher levels of dysfunctional attachment styles, and those high in neuroticism or with a strong desire for being alone overestimated the level of annoyance even more than the average, or underestimated their relationship mood and need satisfaction. Additionally, in Study 2 we examined the effects of factor scores extracted for a latent factor representing general positivity in trait measures. Individuals who scored high on this factor showed less of an overestimating of annoyance, but overestimated their relationship mood.

Finally, our results show that when assessing state relationship satisfaction for more than a few days, the amount of surveys per day seems not to play a crucial role with regard to capturing states representative for the global evaluation of relationship satisfaction. It takes however approximately two weeks to maximize the informational value of the state assessments.

### *Global and Retrospective Assessments of Relationship Satisfaction are Best Represented by the Mean of States (RQ1)*

Our data suggests that when individuals globally or retrospectively evaluate their relationship, they provide information that is foremost reflected by the mean, but also by other summaries of their daily relationship satisfaction states. In contrast to what is described by the peak-and-end rule (Fredrickson, 2000), the 90th quantiles of the state distribution (i.e., positive peaks) and the states reported during the last day explained the lowest amount of variance in retrospective evaluations. Still, recency and peaks represented by the mean of the last week and 10th quantiles (i.e., negative peaks), as well as the median reflected the retrospection only a little bit worse than the mean. Further, descriptively compared, the mean of the first week had lower effects than the mean of the last week; this could support the interpretation of a recency effect during retrospection of relationship satisfaction; but it could also point to individuals developing a certain response pattern over the course of the ESM study, which they draw upon when retrospectively assessing the study period. The development of such a response pattern is supported by the fact that in our longer Study 2 the standard deviation of answers during the first week is significantly higher for all relationship satisfaction items than the standard deviation during the last week (all $p$s < .001). That is, individuals seem to develop a more stable response to the questions, which would undermine the goal of ESM studies to capture state experiences instead of more general beliefs about the relationship. Both interpretations, a recency effect and a more stable response pattern over the course of the ESM study, are possible given the current analyses, and might also both be valid simultaneously.

Our varying results for the different conceptualizations of recency effects (last day, last week) and peaks (highs, lows) are consistent with earlier research: For general daily affect which was retrospectively evaluated on the next day the peak-and-end rule was also not the best explanation, whereas the average of affective states proved to be a good indicator (Miron-Shatz, 2009). The author argues that the end of a day is not special in a sense that some outcome is

reached, which was the case for studies that demonstrated the peak-and-end rule. In the same way were the last days of our study periods not distinctively meaningful for the relationship of our participants. Feldman Barrett (1997) further discusses that the peak-and-end rule was shown for retrospective evaluations that were made immediately after an experience, which was also not the case in our studies (e.g., the mean delay was two days in Study 2).

Regarding incremental effects of other summary statistics beyond the mean, previous research showed for general affect that the lowest (i.e., most negative) affect during a day incrementally explained the retrospective evaluation, whereas the highest (i.e., most positive) affect did not or less so (Ganzach & Yaor, 2018; Miron-Shatz, 2009). This additional effect of intense lows but not highs is plausibly attributed to the general phenomenon of negative experiences weighing more than positive ones (see Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Vaish, Grossmann, & Woodward, 2008 for reviews). Consistent to this, in Study 2, we found that 10th quantiles (i.e., especially negative relationship evaluations) had incremental value to the prediction of retrospection above the effect of the mean of states, for all but the need satisfaction item, whereas the 90th quantiles of the states had an incremental effect only for the retrospection of annoyance (i.e., when individuals were not annoyed at all by their partner). We propose an additional explanation for 10th quantiles providing more information than 90th quantiles: The distribution of relationship satisfaction was skewed in the direction of positive evaluations (most strongly for the annoyance item, mean skew in Study 2 = −3.67). In consequence, 90th quantiles were highly similar to mean values (thereby reducing the informational value compared to 10th quantiles) and had low variance across the sample because of a ceiling effect. Thus, the predictive value the 90th quantiles could provide was limited from the start.

The fact that they still improved the prediction significantly in case of annoyance, might be explained with the observed negative coefficient: The 90th quantile seemingly corrects the error the skewedness introduced to the effect of the mean state. This kind of correction seems to also be provided by the median, as it had also a negative coefficient, being significant for the relationship mood item and the scale of all items. Therefore, characteristics of the distributions of the constructs that are studied must be considered as they might influence which summary statistic improves the prediction.

Finally, even when the mean across all states was already entered in the regression, the average state of the last week and of the last day did still provide significant incremental information for the prediction of the (positively framed) retrospective relationship mood and need satisfaction items, but not for the (negatively framed) annoyance item. Consistent to this result, end evaluations seem to matter more for positive affect than for negative affect (Ganzach & Yaor, 2018).

In sum, our results suggest that the use of the mean as a summary statistic of individuals' relationship satisfaction states is a valid option when the goal is to represent what is captured by retrospective or global evaluations. Vice versa, such global evaluations primarily indicate individuals' average experiences. Still, our data show that especially negative relationship evaluations (e.g., captured by the 10th quantile of a distribution) provide additional information. Exceptionally positive evaluations as indicated by the 90th quantile, or the median might only be incrementally relevant when encountering skewed distributions. Averages of states that are more proximal to the time of retrospection provide in our study an incremental effect for positively framed items. All of these incremental effects may have a functional basis, and may cause a single retrospective assessment to be especially influenced by salient events (see also Lay et al., 2017).

### *Individuals Overestimate their Level of Annoyance in Retrospection (RQ2), which is Moderated by Global Evaluations of the Relationship and Person Characteristics (RQ3)*

#### Overall mean-level bias

When comparing the retrospective relationship satisfaction with the average state during the study period, our data showed significantly different evaluations of the annoyance item, but not of the relationship mood and need satisfaction items. Specifically, individuals overestimated the amount of them having been annoyed by their partner, which results in a lower relationship satisfaction score in retrospection compared to the averaged states (i.e., a negative mean-level bias), if annoyance is included in a scale of relationship satisfaction.

This result cannot be explained by the initial elevation bias found for subjective reports (Shrout et al., 2017), as individuals report an elevated level of annoyance by their partner *after* repeated assessment. It also contrasts the general trend for a positive mean-level bias found in the meta-analysis of Fletcher and Kerr (2010) across judgment categories ("positive" in the sense of evaluating the relationship and the partner better than the relationship or the partner actually is, not in the sense of a general overestimation in retrospection). However, depending on the target of the evaluation, the meta-analysis showed variance in the direction of biases, which is reflected in our results. Previous research which focused on retrospection of relationship experiences found that individuals overestimate their (positively framed) relationship satisfaction, but also their own and their partner's daily positive *and* negative behaviors (Oishi & Sullivan, 2006). This might point to a general pattern of overestimating the occurrence or intensity of specific experiences, independent of the target of evaluation. Miron-Shatz et al. (2009) found such an overestimation trend for general affect (see also Thomas & Diener, 1990; Mitchell, Thompson, Peterson, & Cronk, 1997), but it was stronger for negative affect (see also a recent study by Neubauer et al., 2019 that also shows an overestimation of negative affect in retrospection, but less so for positive affect). It is therefore noteworthy that a) despite referring to our result as a negative mean-level bias (because the relationship quality is described worse in retrospection compared to the averaged state), we observed an

*over*estimation in retrospection, b) this overestimation occurred for the negatively framed domain of annoyance. Negative information dominate positive ones in various domains (see Baumeister et al., 2001; Vaish et al., 2008 for reviews). Lay et al. (2017) argue that the arousal that accompanies an affective reaction is an important factor for the relevance of an experience. Following these ideas, individuals might remember instances of them having been annoyed more profoundly, because these situations were accompanied with negative *and* aroused affect, in contrast to the average positive, not especially aroused daily relationship mood and need satisfaction in healthy relationships.

#### Moderation of mean-level bias by global relationship satisfaction

This line of reasoning is further supported by the fact that global relationship satisfaction showed a clear pattern of moderating the mean-level bias for every item: The unhappier individuals were globally with their relationship, the lower they rated their relationship mood and need satisfaction during the study period (which then was probably more often accompanied with negative emotions), and the higher they rated their level of annoyance in retrospection. Accordingly, the globally happier individuals indicated to be, the closer was their retrospective assessment to the average ESM reports, eventually showing the trend of overestimating the relationship satisfaction in comparison. This result extends findings highlighting the role of global relationship satisfaction for retrospective relationship reports (e.g., Halford, Keefer, & Osgarby, 2002), and its moderating role of bias and accuracy across a range of other judgement categories (Fletcher & Kerr, 2010). Research by Galak and Meyvis (2011) shows that an overestimation of aversive experiences is especially pronounced when individuals expect such experiences in the future. Being annoyed and having one's needs frustrated can be considered aversive experiences. Individuals who are globally unhappy in their relationship have a good reason to expect similar experiences in the future, under the assumption that relationships do not break up easily. From a coping perspective, a study by Luong, Wrzus, Wagner, and Riediger (2016) indicates that valuing negative affect may even be functional with regard to psychosocial and physical functioning. It may therefore be adaptive to focus on negative experiences when remembering the past, to brace for and adapt to similar future relationship episodes.

Compared to an assessment before the ESM study period, global relationship satisfaction concurrently assessed with the retrospection showed the strongest moderating effect. Thus, the recall process seems to be strongly affected by individuals' momentary evaluations, as suggested by Ross (1989), thereby replicating early findings (Holmberg & Holmes, 1994; Karney & Coombs, 2000; McFarland & Ross, 1987). It is important to emphasize that although global relationship satisfaction was quite stable across the four weeks ($r_{CSI}$ = .82 for women and $r_{CSI}$ = .79 for men), the *concurrent* assessments of global relationship satisfaction showed the strongest and most robust effects. That is, the concurrent evaluation of the relationship seems to capture information beyond the stable variance of global relationship satisfaction, which could be interpreted as state variance that is shared with and relied upon during retrospective evaluations (the correlation between retrospection as a scale and the concurrent CSI was $r$ = .70 for women and men). However, studies examining the processes involved when individuals evaluate their global *life* satisfaction find little evidence of experientially induced mood on individuals' evaluations (Yap et al., 2016). Future studies should therefore examine the effect of experientially induced momentary relationship feelings on the recall and global evaluation of relationship satisfaction.

#### Moderation of mean-level bias by other person characteristics

Additional moderating variables support the idea that individuals draw on stable identity-related and situation-specific beliefs when they report on experiences retrospectively (Robinson & Clore, 2002b): Satisfaction with life, which encompasses the belief that one's life is good, had a positive moderating effect (see also Diener et al., 1984), whereas avoidant and anxious attachment styles, which capture negative situation-specific expectations, had negative moderating effects (see also Overall et al., 2015; Pietromonaco & Feldman Barrett, 1997). Similarly, neuroticism moderated the negative mean-level bias of the more affective annoyance item, showing that individuals high in neuroticism overestimate their level of annoyance even stronger. This result mirrors the finding that individuals high in neuroticism overestimate their negative affect in retrospection (Feldman Barrett, 1997), and suggests that this effect generalizes to relationship-specific evaluations as well. Additionally, the explicit desire for closeness had a positive moderating effect on the assessment of the annoyance item, whereas individuals' explicit desire for being alone had a negative moderating effect on the relationship mood item. Previous research already shows that motivational variables influence the recall of autobiographical events (e.g., what experiences are remembered, Woike, 1995; or how the partner behaved, Pusch et al., 2019). It is assumed that during memory retrieval individuals' explicit motives modulate which experiences they capitalize on, namely events that support or were key in changing their self-concept of their goals (Woike, 2008). In this line of reasoning it is sensible that individuals with a strong explicit desire for closeness do not overestimate the level of annoyance as much, as these experiences work against reaching their goal of feeling close to their partner, and are hindering in maintaining a coherent fit between one's goals and one's experiences. In contrast, capitalizing on one's relationship mood when it was bad helps reaffirming the self-concept for individuals who have a strong explicit desire for being alone, that is for individuals who indicate that they regularly need distance from their partner and time for themselves. It is however unclear why only specific items of relationship satisfaction were moderated by the desires, but not others. In sum, rather than giving each experience in their relationship equal meaning during retrospection, individuals seem to capitalize on certain experiences

based on their expectations about the relationship, their impression of themselves and their self-ascribed desires.

As the evaluation of the annoyance item was the main reason for the mean-level bias, and therefore apparently especially susceptible to distortion, we found further moderators that only affected the assessment of this item: In line with the previous moderators, intimacy in the relationship (an indicator of a satisfying relationship with regard to closeness, Laurenceau, Barrett, & Rovine, 2005) had a positive moderating effect for the retrospection of annoyance, reducing the difference between these assessment modalities towards a more similar perception. Surprisingly, the personality factor of conscientiousness turned also out to be a positive moderator. It might be related to a more thorough process when answering the questions, and therefore a more balanced retrospective evaluation as result.

#### Moderation of mean-level bias by a global positivity factor

Given that we found positive moderating effects for constructs that might be perceived as positive (e.g., relationship/life satisfaction), and negative moderating effects for those that might be perceived as negative (e.g., dysfunctional attachment, neuroticism), our results might not be driven by the specific constructs we examined, but alternatively reflect a more general positivity effect or a response style. We considered this possibility by examining a single factor across all self-report items as additional moderator in Study 2: The item loadings suggest that such a factor could be interpreted as a more global identity-related positive self-view about oneself, one's life and one's relationship. Alternatively, it might also reflect a response style characterized by social desirability. This factor indeed moderates the mean-level bias of the annoyance and of the relationship mood item. Hence, depending on the interpretation of the factor, differences between retrospection and the averaged ESM reports seem to be also explained by individuals' global positivity or negativity, or the degree to which they are prone to social desirable responding.

When examining the aforementioned specific moderators simultaneously with this general factor, some moderator effects disappeared, but some other were robust to this control analysis: This suggests that we can confidently interpret some constructs as being relevant as specific moderators of mean level bias. For example, all effects of the relationship satisfaction concurrently assessed with retrospection remained significant, as well as most effects of life satisfaction and relationship satisfaction assessed before the study period. Hence, beyond a general positive assessment of self-report scales, these constructs capture unique variance in satisfaction with specific domains at specific time-points, which explain mean-level differences between retrospection and averaged ESM reports. This robustness was also the case for conscientiousness and anxious attachment as moderators of the annoyance assessment.

The effects of the other moderators (e.g., of avoidant attachment, neuroticism, intimacy, and explicit desires) seem to be more readily explained to be driven by a general positivity/negativity effect. Therefore, our prior interpretations regarding the processes that might cause these specific constructs to moderate the observed differences might be confounded with the effects of a general positive or negative attitude, and should be treated with caution.

#### Summary of moderating effects

In sum, our results suggest that when individuals globally indicate to be unhappy, on average the retrospective reports will suggest a higher occurrence of negative experiences in the relationship as what would be derived from the average of momentary reports. This difference is more pronounced the globally unhappier the individuals are, and is also influenced by aspects of individuals' attachment styles, personality, and global positivity during self-report assessments.

We did neither find effects of gender, as it was found for other judgment domains (Fletcher & Kerr, 2010), nor for delay of retrospection, as would be derived from the accessibility model (Robinson & Clore, 2002a, although we did not systematically vary different delay periods; see Supplemental Materials for estimates of the respective models).

#### Origination of the bias: Retrospection or ESM reports?

In our analyses, we treated the mean ESM measure as truth criterion, with deviations from it during retrospection as bias. This modeling choice has consequences for our interpretation, which have to be carefully considered. First, this assumes that averaging the states is the correct way of summarizing the multiple moments of (dis-) satisfaction an individual experienced during the study, rather than giving the satisfaction during certain situations more weight than other situations (e.g., when spending time with the partner or during a conflict). Second, this modeling of ESM states as the reference criterion might be suggestive of these assessments being not or at least less biased than retrospective assessments. However, while ESM reports might produce fewer recall errors than retrospection, they might be equally or more strongly affected by other response biases, such as those generated by one's self-concept (see Finnigan & Vazire, 2018 for a discussion of such "self-biases" for ESM reports). In fact, we could have modeled the retrospection as truth criterion, with deviations of the aggregated ESM states as bias: This would have led to the interpretation that aggregated ESM reports underestimate the amount of annoyance that "actually" (according to retrospection) occurred in the relationship.

We would like to emphasize that our decision to model the ESM reports as truth criterion impacts the way we interpret our results (i.e., as the retrospective assessment being biased in the sense of an over- or underestimation), but that this choice could reasonably be made differently by other researchers. Importantly, our goal was not to present the ESM reports as the objective gold standard (which was rather a side effect of a modeling decision we had to make), but to uncover any differences between retrospection and aggregated ESM reports. The fact that

these two measures deviate from each other, may be due to different measurements models being applied for representing the relationship satisfaction during the study period, and may lead to the practical implication that the different measurements produce reports with differential validity, which may be useful for different purposes. For example, one could speculate that for couple therapy the retrospective assessment may be more suited to indicate dysfunctional recall biases, and the need of interventions aimed at cognitive reframing, while the aggregation of momentary assessments may draw attention to the influence of situations which might be otherwise less salient.

### *A Saturation Effect is Visible after Assessing Relationship Satisfaction States for Two Weeks (RQ4)*

We also investigated what informational value different sampling schemes of ESM assessments provide with regard to capturing a global assessment of relationship satisfaction. We examined two factors that can be manipulated when designing an ESM study: The number and the scheduling of the assessments.

The number of assessments can be influenced in two ways: By increasing the number of assessments per day, or by increasing the overall length of the study. Both ways of collecting more experiences have pros and cons (e.g., capturing short-term dynamics vs. enhancing participant burden) and must be decided depending on the research question at hand (see Bolger & Laurenceau, 2013). The decisions are however not independent, as a less intensive sampling per day may invoke the need for a longer study period to achieve representative information. In our data it takes about five days to achieve a similar overall level of association with global relationship satisfaction, regardless of whether only one random sample per day is considered or five semi-random samples per day. After five days, the increase in association strength is similar steady across different numbers of assessments per day, maxing at around $\beta = .60$ (but see Schönbrodt et al., 2019 demonstrating high within-day variance of state relationship satisfaction, which raises the need to sample multiple times per day to capture the dynamics occurring within a day). Further, we see a saturation effect after approximately two weeks, meaning that after this study period more ESM data does not provide much more incremental information for predicting global relationship satisfaction – independently of the number of assessments per day. This complements the findings of Epstein (1979), who also found two weeks to be necessary for achieving a representative sample of individual's behaviors.

Regarding the timing of the assessments, we examined three common strategies: Assessing in the evening, in the morning, or at a random time during the day. While we descriptively found in our larger Study 2 that evening assessments seem to be more valid for representing global relationship satisfaction, because both the initial association strength was higher and the maximum association strength was reached sooner, this did not replicate in our smaller Study 1. Hence, further research is needed to assess the robustness of the differences between sampling plans when only sampling once.

### *Limitations*

Several potential limitations have to be considered when interpreting the results of our studies. First, a necessary condition for the investigation of bias and accuracy (RQ2 and RQ3) is the commensurability of the measures that are being compared, in our case of the retrospection and the state assessment. In principle, this is given in the current studies, as the same content is evaluated in both measures (leading to "nominal equivalence")[16] on the same scale (transformed to the same metric, leading to "scale equivalence"; see Edwards & Shipp, 2007 for the use of these terms). However, slightly different assessment characteristics for ESM and retrospection, especially visual differences in the presentation of the sliders used, could pose a threat to commensurability: The retrospective assessment was answered in a browser on the participants' personal computers, and in Study 2 the three relationship satisfaction items were presented in a block. The ESM assessment, in contrast, was completed on the smartphone and the items were presented at different positions in the ESM survey (but see Wells, Bailey, & Link, 2014, finding little psychometric differences between web and smartphone presentation of items). Further, slightly different slider characteristics might have biased the answers (see Matejka, Glueck, Grossman, & Fitzmaurice, 2016). First, a missing "neutral" label in the retrospective assessment could have removed an anchor effect that might have been present in ESM. However, the largest biases were found for the annoyance item, which also in the ESM assessment did not have a neutral label (see **Figure 1**). Second, the slider having a start position during retrospection, whereas in ESM no start value was preselected, could have evoked another anchoring effect. As the start position was in the middle of the scale, this might have canceled out the missing "neutral" option for the relationship mood and need satisfaction items. For the annoyance item this might actually have introduced a biased anchoring point, although it is unclear why this would produce an overestimation of annoyance: Participants rather seem to choose preselected options less often (Funke, 2016), that is, the preselection seems to evoke the need to move the slider further away; given that on an absolute level the amount of annoyance reported was low (mean of retrospection of not reverse scored annoyance = 1.72 on a scale from 0 to 10) and the labeled end of the scale "*not at all* (*annoyed*)" might attract answers, these kind of biasing design effects should have rather led to an underestimation of annoyance, rather than the observed overestimation. Finally, although we transformed all measures to the same metric (0–10), the ESM answers on the slider items were initially saved in a higher resolution (on scales from 1–7 and 0–10 with answers saved with multiple positions after the decimal point) than the retrospective evaluations (on scales from 0–100 and 1–100 with answers rounded to whole numbers). To assess the magnitude of error these different resolutions might have added to our results, we adjusted the resolution of the ESM answers in Study 2 to the answers during retrospection by transforming them to a 1–100 scale, rounding them to whole numbers, and

transforming them back to a 0–10 scale. All of the results replicate when running the analyses with these scales, with changes in the estimates only on the third or fourth decimal place after the comma.

Further, our analyses showed that a mean-level bias primarily occurs for the retrospection of experienced annoyance, therefore biasing the whole relationship satisfaction scale in retrospection when this item is included in scale calculation. Therefore, our results may not generalize for other relationship satisfaction scales that do not include annoyance, or maybe more generally those scales that do not contain items pertaining to negative affect in the relationship. We would argue, however, that simply removing the annoyance item, or more generally avoiding the assessment of negative affectivity in relationships is no solution. As also discussed in Schönbrodt et al. (2019), the annoyance item contributes to a more heterogeneous index of relationship satisfaction, taking into account the impact of negative experiences for relationship evaluation (as other scales also do, e.g., the global measures applied in our studies, Funk & Rogge, 2007; Rogge et al., 2016). Depending on the research question, this broader assessment of relationship satisfaction is necessary to achieve a complete picture of individuals' relationship evaluation and may be more suited to differentiate couples in generally happy relationships.

Moreover, our analyses concerning the required number of ESM surveys and the optimal sampling procedure to reach satisfactory associations with a global evaluation were not based on an experimental design: All participants answered the same amount of five surveys with a semi-random schedule, but for our analyses we selected different subsets of surveys as predictors of global relationship satisfaction. In consequence, the effects we found might differ if individuals would actually only answer one survey (or fewer than five surveys) per day (in the morning or in the evening), as the ESM procedure we applied could have induced reactivity such as a heightened sensitivity for participant's relationship feelings. If this would be the case, then our effects might be exaggerated, and a lower number of surveys for instance might take longer than the reported five days to reach a similar association strength as a higher number of surveys. Future work should compare the effects we found in our study with effects from an experimental study which randomly assigns participants to different ESM designs.

Finally, despite the fact that we preregistered some hypotheses for RQ2, the presented results should mainly be regarded as exploratory, as we were inconsistent in the preregistration regarding which items we will use as a measure of state and retrospective relationship satisfaction. For maximal transparency and given the exploratory nature of the other research questions, we reported the results for all available items, and controlled the false discovery rate at $\alpha = 5\%$.

## Conclusion

The present studies provide insight into various domains related to the assessment of relationship satisfaction. First, our studies showed that global and retrospective evaluations best capture the average of relationship satisfaction states, with other summary statistics providing incremental information. Second, the retrospective overestimation of negative affect found in prior research also holds for a relationship-specific negative evaluation of annoyance. Third, this difference between retrospective and aggregated ESM assessments is especially pronounced for individuals who globally report low relationship and life satisfaction, with other person characteristics being further relevant. Last, our results show that approximately two weeks are necessary to sample a representative amount of relationship satisfaction states. The current research uncovers differences of various assessment modalities of relationship satisfaction that ought to be considered when applying them: Retrospective assessments and in extension also global evaluations might provide notably different information than aggregated ESM reports when targeting negative experiences in a relationship, especially for individuals who globally report to be unhappy. Depending on the research question or the aim of assessment in a practitioner setting, it has to be carefully decided whether one is interested in the average of the experiences that were reported to happen in the relationship, with each of these momentary reports probably having their own biases; or whether the idiosyncratic capitalization individuals make for specific experiences is of special interest, which is provided by retrospective or global measures.

## Data Accessibility Statement

The data of both studies are available as a scientific use file (Zygar et al., 2018b for Study 1; Zygar-Hoffmann, Hagemeyer, Pusch, & Schönbrodt, 2020 for Study 2).

We embrace the values of openness and transparency in science (http://www.researchtransparency.org/). We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study (Simmons, Nelson, & Simonsohn, 2012). The preregistration of Study 1 can be found at https://osf.io/hafsx/, the preregistration of Study 2 at https://osf.io/af4yb/. Both preregistrations contain additional hypotheses on other research questions than the ones reported here. The Supplemental Materials for this paper and complete codebooks can be found at https://osf.io/sq7mw/. The codebooks include all variables of the studies, also those not included in the current paper.

## Notes

[1] We did preregister in both studies that the mean of relationship satisfaction is positively related to retrospection (see tracking accuracy hypothesis described for RQ2) as well as to global relationship satisfaction (see p. 9 and p. 41). However, our main goal in RQ1 was to descriptively compare the different summary statistics for the prediction of retrospection and global relationship satisfaction, but our preregistrations do not mention other summary statistics than the mean. Hence, even though the preregistered hypotheses also correspond to two analyses reported for RQ1, we

refrain to draw special attention to these analyses being preregistered.

[2] Zygar et al. (2018a) also report the result of regressing the global relationship satisfaction evaluation on mean relationship satisfaction states (corresponding to one single coefficient of **Table 3**), but in that paper the ESM states were *z*-standardized *before* aggregating them within each person, thereby the result is not equal to the standardized regression coefficient reported in the current manuscript.

[3] The preregistration contains erroneous time-frames on this matter.

[4] This number is slightly lower than the one reported in Zygar et al. (2018a), because in that paper we reported the number of measurement points before survey-level exclusions ($n = 53$) and included started surveys without a single answered item ($n = 116$).

[5] For one couple, we observed an inconsistency in the gender both partners indicated in the pre-ESM compared to the follow-up-questionnaire one year later. We did not exclude this couple from our main analyses, but as the inconsistency might point to careless responding, we report in the Supplemental Materials how minor results for RQ3 change when excluding this couple from the analyses. For all other RQs the pattern of results does not change when excluding their data.

[6] In our preregistration we erroneously stated that less than 33% of 140 would be less than 24 rather than the actual 47 surveys (see p. 59).

[7] For analyses with data from retrospection, we describe if the pattern of results changes when not excluding this data.

[8] In total, 12 individuals had a Cook's Distance of >2 *SD*. However, two of these individuals were not treated as outliers, as they were just very unhappy with their relationship (and thus different than the majority of the sample), but still consistent in their answers, in contrast to the other 10 individuals who indicated high positive relationship mood and need satisfaction, but also high annoyance.

[9] In our preregistrations we stated we would use two-intercept models as default (i.e., separate intercepts for men and women, see p. 18 and p. 5). However, in the current case, using a gender contrast variable leads to a more meaningful interpretation of the intercept (mean-level bias across both genders, instead of a mean-level bias separately for men and women).

[10] In the preregistration of Study 2 we mention "For controlling the false-discovery-rate (FDR) at 5% we will apply the Benjamini-Hochberg procedure [...]" (p. 6), but we also state that "For exploratory analyses, we consider effects noteworthy when $p < .01$ and $\beta \geq .05$ (for additional moderations of hypotheses) or $\beta \geq .10$ (for additional main effects)" (p. 6). Both procedures lead to reporting roughly equivalent exploratory effects in the current paper. We decided on the FDR procedure, as the number of effects to control for could be determined (number of analyses = number of summary statistics or moderators multiplied by the number of items plus the scale; separately for mean-level bias and tracking accuracy) and the other procedure can more easily be applied by the readers themselves.

[11] We also explored the results for regressing the retrospective assessment on the median, the mean of the last week, of the last day, and of the first week, controlling the FDR for the according number of tests. The reported mean-level bias for the annoyance item and the scale replicated for all of these summary statistics, and even extended to the other two items in some cases.

[12] Therefore, this result could have also been labeled as a positive mean-level bias of annoyance, in the sense of an overestimating of the variable of interest. However, to avoid confusion and to consistently refer to "negative mean-level bias" as assessing the relationship in retrospection worse than what was indicated by the ESM reports, we label the difference that occurred in retrospection of (reverse-scored) annoyance as negative mean-level bias as well (see Fletcher & Kerr, 2010, who also use these terms accordingly).

[13] As all relationship satisfaction variables had skewed distributions, all of our models had an overall poor fit. We reran the analyses of RQ2 as Bayesian MLMs in the *brms* package (Bürkner, 2017) with default priors, but specifying skewed normal distributions with an inverse and a log link. These alternative models fitted better, although still not good in case of annoyance. The results were consistent with those reported here. When specifying a log link with the skewed normal distribution, additionally a negative mean-level bias for the relationship mood and the need satisfaction item emerged (with the 95%-HDI of the intercepts excluding zero).

[14] When not excluding low quality responses (see Sample) the tracking accuracy of the scale is no longer significantly different from one.

[15] When not excluding low quality responses (see Sample) the moderation for the annoyance item by life satisfaction, avoidant attachment, neuroticism, and conscientiousness was no longer significant. PRQ assessed before ESM and anxious attachment are no longer significant moderators for any item. Instead, a significant moderation by gender for the need satisfaction item indicates an underestimation by men and an overestimation by women, and self-reflection shows a significant positive moderation for the annoyance item.

[16] A study by Winkielman, Knäuper, and Schwarz (1998) suggests that when referring to different time frames in questionnaires, the interpretation of the phenomenon that is being assessed changes. Specifically, the study provides evidence that a reference to longer time frames (e.g. "during the last month") prompt individuals to report less frequent, but more intense events, compared to a reference to shorter time frames (e.g. "during the last week"). The authors explain this with the ambiguity of the phenomena that are studied, and note that an explicit definition of the phenomenon resolves

this problem; importantly, they also show that the interpretation elicited by a reference to a shorter time frame carries over when subsequently a longer time frame is assessed (although this did not completely eliminate the effect of the time frame, at least not for frequency reports). Such a carry-over effect is to be expected in our study, as individuals could internalize the meaning of the different relationship satisfaction items multiple times per day for several weeks. Although we cannot rule out that their interpretation of the relationship satisfaction items changed when they were asked to assess them retrospectively for the study period right after the study, we do not find it plausible that they did not recognize the questions and interpreted the item content differently as during the multiple instances they assessed it during the prior weeks.

## Acknowledgements

The authors thank Birk Hagemeyer, Sebastian Pusch, Paula Fehrmann, Nicole Horn, and Lara Lietge for the PACT scoring, Helen Baumann for assistance during data collection of Study 1, as well as Tobias Kächele, Lukas Müller and Ludwig Zellner for the app development.

## Funding Information

This research was funded by grants from the German Research Foundation to Felix Schönbrodt (SCHO 1334/5-1) and Birk Hagemeyer (HA 6884/2-1).

## Competing Interests

The authors have no competing interests to declare.

## Author Contributions

- Contributed to conception and design: CZ-H, FS
- Contributed to acquisition of data: CZ-H
- Contributed to analysis and interpretation of data: CZ-H, FS
- Drafted and/or revised the article: CZ-H
- Approved the submitted version for publication: CZ-H, FS

## References

**Arslan, R.,** & **Tata, C.** (2016). *formr.org survey software (version v0.16.12)*. Retrieved from https://formr.org/

**Arslan, R.,** & **Tata, C.** (2017). *formr.org survey software (version v0.16.13–version v0.17.16)*. Retrieved from https://formr.org/

**Arslan, R. C., Walther, M. P.,** & **Tata, C. S.** (2018). Formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods*, 1–12. DOI: https://doi.org/10.31234/osf.io/pjasu

**Aust, F.,** & **Barth, M.** (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

**Barton, K.** (2018). *MuMIn: Multi-model inference*. Retrieved from https://CRAN.R-project.org/package=MuMIn

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

**Baumeister, R. F., Bratslavsky, E., Finkenauer, C.,** & **Vohs, K. D.** (2001). Bad is atronger than good. *Review of General Psychology*, *5*(4). DOI: https://doi.org/10.1037/1089-2680.5.4.323

**Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., ... Wrzus, C.** (2017). Integrating personality structure, personality process, and personality development. *European Journal of Personality*, *31*(5), 503–528. DOI: https://doi.org/10.1002/per.2115

**Bell, K. J.** (2008). *Intimate partner violence on campus: A Test of social learning theory* (PhD thesis). Retrieved from https://pdfs.semanticscholar.org/730b/49e22394259ac4a1c96692f9ec8c280ccff4.pdf

**Benjamini, Y.,** & **Hochberg, Y.** (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. DOI: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

**Biderman, M. D., Nguyen, N. T., Cunningham, C. J.,** & **Ghorbani, N.** (2011). The ubiquity of common method variance: The case of the big five. *Journal of Research in Personality*, *45*(5), 417–429. DOI: https://doi.org/10.1016/j.jrp.2011.05.001

**Blood, R.,** & **Wolfe, D.** (1960). *Husbands and wives*. Glencoe, IL: Free Press.

**Bolger, N.,** & **Laurenceau, J.-P.** (2013). *Intensive longitudinal methods*. New York, NY: Guilford.

**Bürkner, P.-C.** (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. DOI: https://doi.org/10.18637/jss.v080.i01

**Chang, V. T., Overall, N. C., Madden, H.,** & **Low, R. S.** (2018). Expressive suppression tendencies, projection bias in memory of negative emotions, and well-being. *Emotion*, *18*(7), 925–941. DOI: https://doi.org/10.1037/emo0000405

**Christensen, T. C., Wood, J. V.,** & **Feldman Barrett, L.** (2003). Remembering everyday experience through the prism of self-esteem. *Personality and Social Psychology Bulletin*, *29*(1), 51–62. DOI: https://doi.org/10.1177/0146167202238371

**Conner, T. S.,** & **Feldman Barrett, L.** (2012). Trends in ambulatory self-report: The role of momentary experience in psychosomatic medicine. *Psychosomatic Medicine*, *74*, 327–337. DOI: 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. DOI: https://doi.org/10.1097/00005053-198709000-00004

**Dewey, M.** (2018). *metap: Meta-analysis of significance values*. Retrieved from https://cran.r-project.org/package=metap

**Diener, E., Emmons, R. A., Larsen, R. J.,** & **Griffin, S.** (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*(1), 71–75. DOI: https://doi.org/10.1207/s15327752jpa4901_13

**Diener, E., Larsen, R. J.,** & **Emmons, R. A.** (1984). Bias in mood recall in happy and unhappy persons. *Paper presented at the 92nd Annual Convention of the American Psychological Association*, Toronto, Ontario, Canada.

**Edwards, J. R.,** & **Shipp, A. J.** (2007). The relationship between person-environment fit and outcomes: An integrative theoretical framework. In C. Ostroff & T. A. Judge (Eds.), *Perspectives on organizational fit* (pp. 209–258). San Francisco: Jossey-Bass.

**Ehrenthal, J. C., Dinger, U., Lamla, A., Funken, B.,** & **Schauenburg, H.** (2009). Evaluation der deutschsprachigen Version des Bindungsfragebogens "Experiences in Close Relationships – Revised" (ECR-RD) [Evaluation of the German Version of the Attachment Questionnaire "Experiences in Close Relationships – Revised" (ECR-RD)]. *PPmP-Psychotherapie Psychosomatik Medizinische Psychologie*, *59*(6), 215–223. DOI: https://doi.org/10.1055/s-2008-1067425

**Epstein, S.** (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, *37*(July). DOI: https://doi.org/10.1037/0022-3514.37.7.1097

**Fallis, E. E., Rehman, U. S., Woody, E. Z.,** & **Purdon, C.** (2016). The longitudinal association of relationship satisfaction and sexual satisfaction in long-term relationships. *Journal of Family Psychology*, *30*(7), 822–831. DOI: https://doi.org/10.1037/fam0000205

**Feldman Barrett, L.** (1997). The relationship among momentary emotianl experiences, personaltiy descriptions, and retrospective ratings of emotion. *Personality and Social Psychology Bulletin*, *23*, 1100–1110. DOI: https://doi.org/10.1177/01461672972310010

**Fincham, F. D.,** & **Rogge, R. D.** (2010). Understanding relationship quality: Theoretical challenges and new tools for assessment. *Journal of Family Theory & Review*, *2*(4), 227–242. DOI: https://doi.org/10.1111/j.1756-2589.2010.00059.x

**Finnigan, K. M.,** & **Vazire, S.** (2018). The incremental validity of average state self-reports over global self-reports of personality. *Journal of Personality and Social Psychology*, *115*(2), 321–337. DOI: https://doi.org/10.1037/pspp0000136

**Fleeson, W.** (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011–1027. DOI: https://doi.org/10.1037/0022-3514.80.6.1011

**Fleeson, W.,** & **Gallagher, P.** (2009). The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of Personality and Social Psychology*, *97*(6), 1097–1114. DOI: https://doi.org/10.1037/a0016786

**Fletcher, G. J.,** & **Kerr, P. S.** (2010). Through the eyes of love: Reality and illusion in intimate relationships. *Psychological Bulletin*, *136*(4), 627–658. DOI: https://doi.org/10.1037/a0019792

**Forbes, E. E., Stepp, S. D., Dahl, R. E., Ryan, N. D., Whalen, D., Axelson, D. A., … Silk, J. S.** (2012). Real-world affect and social context as predictors of treatment response in child and adolescent depression and anxiety: An ecological momentary assessment study. *Journal of Child and Adolescent Psychopharmacology*, *22*(1), 37–47. DOI: https://doi.org/10.1089/cap.2011.0085

**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. DOI: https://doi.org/10.1080/026999300402808

**Funk, J. L.,** & **Rogge, R. D.** (2007). Testing the ruler with item response theory: Increasing precision of measurement for relationship satisfaction with the Couples Satisfaction Index. *Journal of Family Psychology*, *21*(4), 572–583. DOI: https://doi.org/10.1037/0893-3200.21.4.572

**Funke, F.** (2016). A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales. *Social Science Computer Review*, *34*(2), 244–254. DOI: https://doi.org/10.1177/0894439315575477

**Galak, J.,** & **Meyvis, T.** (2011). The pain was greater if it will happen again: The effect of anticipated continuation on retrospective discomfort. *Journal of Experimental Psychology: General*, *140*(1), 63–75. DOI: https://doi.org/10.1037/a0021447

**Ganzach, Y.,** & **Yaor, E.** (2018). The retrospective evaluation of positive and negative affect. *Personality and Social Psychology Bulletin*, *45*(1), 93–104. DOI: https://doi.org/10.1177/0146167218780695

**Gerlitz, J.-Y.,** & **Schupp, J.** (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. [Collection of the Big-Five personality characteristics in SOEP]. *DIW Research Notes*, *4*, 1–36.

**Glaesmer, H., Grande, G., Braehler, E.,** & **Roth, M.** (2011). The German version of the Satisfaction With Life Scale (SWLS). *European Journal of Psychological Assessment*, *27*(2), 127–132. DOI: https://doi.org/10.1027/1015-5759/a000058

**Gordon, M.,** & **Lumley, T.** (2017). *Forestplot: Advanced forest plot using 'grid' graphics*. Retrieved from https://CRAN.R-project.org/package=forestplot

**Grant, A., Franklin, J.,** & **Langford, P.** (2002). The Self-Reflection and Insight Scale: A new measure of private self-consciousness. *Social Behavior and Personality: An International Journal*, *30*(8). DOI: https://doi.org/10.2224/sbp.2002.30.8.821

**Greischel, H.,** & **Johnson, M.** (n.d.). *Measurement invariance of a German translation of the Couples Satisfaction Index*.

**Hagemeyer, B.,** & **Neyer, F. J.** (2012). Assessing implicit motivational orientations in couple relationships: The Partner-Related Agency and Communion Test (PACT). *Psychological Assessment*, *24*(1), 114–28. DOI: https://doi.org/10.1037/a0024822

**Hagemeyer, B., Neyer, F. J., Neberich, W.,** & **Asendorpf, J. B.** (2013). The ABC of Social Desires: Affiliation,

being alone, and closeness to partner. *European Journal of Personality*, *27*(5), 442–457. DOI: https://doi.org/10.1037/t33449-000

**Halford, W. K., Keefer, E.,** & **Osgarby, S. M.** (2002). "How has the week been for you two?" Relationship satisfaction and hindsight memory biases in couples' reports of relationship events. *Cognitive Therapy and Research*, *26*(6), 759–773. DOI: https://doi.org/10.1023/A:1021289400436

**Hedges, S. M., Jandorf, L.,** & **Stone, A. A.** (1985). Meaning of daily mood assessments. *Journal of Personality and Social Psychology*, *48*(2), 428–434. DOI: https://doi.org/10.1037/0022-3514.48.2.428

**Hofmann, W., Finkel, E. J.,** & **Fitzsimons, G. M.** (2015). Close relationships and self-regulation: How relationship satisfaction facilitates momentary goal pursuit. *Journal of Personality and Social Psychology*, *109*(3), 434–452. DOI: https://doi.org/10.1037/pspi0000020

**Holmberg, D.,** & **Holmes, J. G.** (1994). Reconstruction of relationship memories: A mental models approach. In N. Schwarz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective reports* (pp. 267–288). DOI: https://doi.org/10.1007/978-1-4612-2624-6_18

**Johnson, P. C. D.** (2014). Extension of Nakagawa and Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946. DOI: https://doi.org/10.1111/2041-210X.12225

**Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N.,** & **Stone, A. A.** (2014). A survey method for characterizing daily life experience: The Day Reconstruction Method. *Science*, *306*(5702), 1776–1780. DOI: https://doi.org/10.1126/science.1103572

**Karney, B. R.,** & **Coombs, R. H.** (2000). Memory bias in long term close relationships: Consistency or improvement? *Personality and Social Psychology Bulletin*, *26*(8), 959–970. DOI: https://doi.org/10.1177/01461672002610006

**Karney, B. R.,** & **Frye, N. E.** (2002). "But we've been getting better lately": Comparing prospective and retrospective views of relationship development. *Journal of Personality and Social Psychology*, *82*(2), 222–238. DOI: https://doi.org/10.1037/0022-3514.82.2.222

**Knowles, J. E.,** & **Frederick, C.** (2018). *MerTools: Tools for analyzing mixed effect regression models*. Retrieved from https://CRAN.R-project.org/package=merTools

**Kuznetsova, A., Brockhoff, P. B.,** & **Christensen, R. H. B.** (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. DOI: https://doi.org/10.18637/jss.v082.i13

**Laurenceau, J.-P., Barrett, L. F.,** & **Rovine, M. J.** (2005). The interpersonal process model of intimacy in marriage: A daily-diary and multilevel modeling approach. *Journal of Family Psychology*, *19*(2), 314–323. DOI: https://doi.org/10.1037/0893-3200.19.2.314

**Lay, J. C., Gerstorf, D., Scott, S. B., Pauly, T.,** & **Hoppmann, C. A.** (2017). Neuroticism and extraversion magnify discrepancies between retrospective and concurrent affect reports. *Journal of Personality*, *85*(6), 817–829. DOI: https://doi.org/10.1111/jopy.12290

**LimesurveyGmbH.** (2017). *LimeSurvey: An open source survey tool.* , Hamburg, Germany: LimeSurvey GmbH. Retrieved from http://www.limesurvey.org

**Lucas, R. E., Wallsworth, C., Anusic, I.,** & **Donnellan, M. B.** (2019). *A direct comparison of the Day Reconstruction Method and the Experience Sampling Method*. DOI: https://doi.org/10.31234/osf.io/cv73u

**Luchies, L. B., Rusbult, C. E., Eastwick, P. W., Wieselquist, J., Kumashiro, M., Coolsen, M. K.,** & **Finkel, E. J.** (2013). Trust and biased memory of transgressions in romantic relationships. *Journal of Personality and Social Psychology*, *104*(4), 673–694. DOI: https://doi.org/10.1037/a0031054

**Luong, G., Wrzus, C., Wagner, G. G.,** & **Riediger, M.** (2016). When bad moods may not be so bad: Valuing negative affect is associated with weakened affect-health links. *Emotion*, *16*(3), 387–401. DOI: https://doi.org/10.1037/emo0000132

**Matejka, J., Glueck, M., Grossman, T.,** & **Fitzmaurice, G.** (2016). The effect of visual appearance on the performance of continuous sliders and visual analogue scales. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 5421–5432. DOI: https://doi.org/10.1145/2858036.2858063

**McCrae, R. R., Bond, M. H., Yik, M. S., Trapnell, P. D.,** & **Paulhus, D. L.** (1998). Interpreting personality profiles across cultures: Bilingual, acculturation, and peer rating studies of Chinese undergraduates. *Journal of Personality and Social Psychology*, *74*(4), 1041–1055. DOI: https://doi.org/10.1037/0022-3514.74.4.1041

**McFarland, C.,** & **Ross, M.** (1987). The relation between current impressions and memories of self and dating partners. *Personality and Social Psychology Bulletin*, *13*(2), 228–238. DOI: https://doi.org/10.1177/0146167287132008

**Mill, A., Realo, A.,** & **Allik, J.** (2016). Retrospective ratings of emotions: The effects of age, daily tiredness, and personality. *Frontiers in Psychology*, *6*(JAN), 1–12. DOI: https://doi.org/10.3389/fpsyg.2015.02020

**Miron-Shatz, T.** (2009). Evaluating multiepisode events: Boundary conditions for the peak-end rule. *Emotion*, *9*(2), 206–213. DOI: https://doi.org/10.1037/a0015295

**Miron-Shatz, T., Stone, A.,** & **Kahneman, D.** (2009). Memories of yesterday's emotions: Does the valence of experience affect the memory-experience gap? *Emotion*, *9*(6), 885–891. DOI: https://doi.org/10.1037/a0017823

**Mitchell, T. R., Thompson, L., Peterson, E.,** & **Cronk, R.** (1997). Temporal adjustments in the evaluation of events: The "Rosy View". *Journal of Experimental Social Psychology*, *33*(4), 421–448. DOI: https://doi.org/10.1006/jesp.1997.1333

**Nakagawa, S.,** & **Schielzeth, H.** (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. DOI: https://doi.org/10.1111/j.2041-210x.2012.00261.x

**Neubauer, A. B., Scott, S. B., Sliwinski, M. J.,** & **Smyth, J. M.** (2019). 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*. DOI: https://doi.org/10.1037/pspp0000248

**Oishi, S.,** & **Sullivan, H. W.** (2006). The predictive value of daily vs. retrospective well-being judgments in relationship stability. *Journal of Experimental Social Psychology*, *42*(4), 460–470. DOI: https://doi.org/10.1016/j.jesp.2005.07.001

**Overall, N. C., Fletcher, G. J., Simpson, J. A.,** & **Fillo, J.** (2015). Attachment insecurity, biased perceptions of romantic partners' negative emotions, and hostile relationship behavior. *Journal of Personality and Social Psychology*, *108*(5), 730–749. DOI: https://doi.org/10.1037/a0038987

**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. DOI: https://doi.org/10.1177/0146167295214003

**Pietromonaco, P. R.,** & **Feldman Barrett, L.** (1997). Working models of attachment and daily social interactions. *Journal of Personality and Social Psychology*, *73*(6), 1409–1423. DOI: https://doi.org/10.1037/0022-3514.73.6.1409

**Pusch, S., Schönbrodt, F. D., Zygar-Hoffmann, C.,** & **Hagemeyer, B.** (2019). Truth and wishful thinking: How inter-individual differences in communal motives manifest in momentary partner perceptions. *European Journal of Personality*. DOI: https://doi.org/10.1002/per.2227

**Rammstedt, B.,** & **John, O. P.** (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212. DOI: https://doi.org/10.1016/j.jrp.2006.02.001

**R Core Team.** (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

**Reis, H. T.,** & **Gable, S. L.** (2000). Event Sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in personality psychology* (pp. 190–222). Cambridge, UK: Cambridge University Press.

**Reise, S. P.** (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. DOI: https://doi.org/10.1080/00273171.2012.715555

**Robinson, M. D.,** & **Clore, G. L.** (2002a). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, *128*(6), 934–960. DOI: https://doi.org/10.1037/0033-2909.128.6.934

**Robinson, M. D.,** & **Clore, G. L.** (2002b). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology*, *83*(1), 198–215. DOI: https://doi.org/10.1037/0022-3514.83.1.198

**Robinson, M. D., Johnson, J. T.,** & **Shields, S. A.** (1998). The gender heuristic and the database: Factors affecting the perception of gender-related differences in the experience and display of emotions. *Basic and Applied Social Psychology*, *20*(3), 206–219. DOI: https://doi.org/10.1207/s15324834basp2003_3

**Rogge, R. D., Fincham, F. D., Crasta, D.,** & **Maniaci, M. R.** (2016). Positive and negative evaluation of relationships: Development and validation of the Positive – Negative Relationship Quality (PN-RQ) Scale. *Psychological Assessment*, *29*(8), 1028–1043. DOI: https://doi.org/10.1037/pas0000392

**Rosenberg, E. L.** (1998). Levels of analysis and the organization of affect. *Review of General Psychology*, *2*(3), 247–270. DOI: https://doi.org/10.1037/1089-2680.2.3.247

**Ross, M.** (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review*, *96*(2), 341–357. DOI: https://doi.org/10.1037/0033-295X.96.2.341

**Rosseel, Y.** (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/. DOI: https://doi.org/10.18637/jss.v048.i02

**Schimmack, U.,** & **Hartmann, K.** (1997). Individual differences in the memory representation of emotional episodes: Exploring the cognitive processes in repression. *Journal of Personality and Social Psychology*, *73*(5), 1064–1079. DOI: https://doi.org/10.1037/0022-3514.73.5.1064

**Schönbrodt, F. D.,** & **Gerstenberg, F. X. R.** (2012). An IRT analysis of motive questionnaires: The Unified Motive Scales. *Journal of Research in Personality*, *46*, 725–742. DOI: https://doi.org/10.1016/j.jrp.2012.08.010

**Schönbrodt, F. D., Zygar-Hoffmann, C., Nestler, S., Pusch, S.,** & **Hagemeyer, B.** (2019). *Measuring motivational relationship processes in experience sampling: A reliability model for moments, days, and persons nested in couples*. DOI: https://doi.org/10.31234/osf.io/6mq7t

**Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., ... Bolger, N.** (2017). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences*, *115*(1), E15–E23. DOI: https://doi.org/10.1073/pnas.1712277115

**Sprecher, S.** (1999). "I love you more today than yesterday": Romantic partners' perceptions of changes in love and related affect over time. *Journal of Personality and Social Psychology*, *76*(1), 46–53. DOI: https://doi.org/10.1037/0022-3514.76.1.46

**Thomas, D. L.,** & **Diener, E.** (1990). Memory accuracy in the recall of emotions. *Journal of Personality and Social Psychology*, *59*(2), 291–297. DOI: https://doi.org/10.1037/0022-3514.59.2.291

**Vaish, A., Grossmann, T.,** & **Woodward, A.** (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383–403. DOI: https://doi.org/10.1037/0033-2909.134.3.383

**Walentynowicz, M., Schneider, S.,** & **Stone, A. A.** (2018). The effects of time frames on self-report. *PLOS ONE*, *13*(8), 1–18. DOI: https://doi.org/10.1371/journal.pone.0201655

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

**Wells, T., Bailey, J. T.,** & **Link, M. W.** (2014). Comparison of smartphone and online computer survey administration. *Social Science Computer Review*, *32*(2), 238–255. DOI: https://doi.org/10.1177/0894439313505829

**West, T. V.,** & **Kenny, D. A.** (2011). The Truth and Bias Model of Judgment. *Psychological Review*, *118*(2), 357–378. DOI: https://doi.org/10.1037/a0022936

**Wickham, H., François, R., Henry, L.,** & **Müller, K.** (2018). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

**Winkielman, P., Knäuper, B.,** & **Schwarz, N.** (1998). Looking back at anger: Reference periods change the interpretation of emotion frequency questions. *Journal of Personality and Social Psychology*, *75*(3), 719–728. DOI: https://doi.org/10.1037/0022-3514.75.3.719

**Woike, B. A.** (1995). Most-memorable experiences: Evidence for a link between implicit and explicit motives and social cognitive processes in everyday life. *Journal of Personality and Social Psychology*, *68*(6), 1081–1091. DOI: https://doi.org/10.1037/0022-3514.68.6.1081

**Woike, B. A.** (2008). A functional framework for the influence of implicit and explicit motives on autobiographical memory. *Personality and Social Psychology Review*, *12*(2), 99–117. DOI: https://doi.org/10.1177/1088868308315701

**Yap, S. C. Y., Wortman, J., Anusic, I., Baker, S. G., Scherer, L. D., Donnellan, M. B.,** & **Lucas, R. E.** (2016). The effect of mood on judgments of subjective well-being: Nine tests of the judgment model. *Journal of Personality and Social Psychology*, *113*(6), 939–961. DOI: https://doi.org/10.31234/osf.io/5fj9c

**Zygar, C., Hagemeyer, B., Pusch, S.,** & **Schönbrodt, F. D.** (2018a). From motive dispositions to states to outcomes: An intensive experience sampling study on communal motivational dynamics in couples. *European Journal of Personality*, *32*(3), 306–324. DOI: https://doi.org/10.1002/per.2145

**Zygar, C., Hagemeyer, B., Pusch, S.,** & **Schönbrodt, F. D.** (2018b). From motive dispositions to states to outcomes: Research data of an intensive experience sampling study on communal motivational dynamics in couples [translated title] (version 2.0.0) [data and documentation]. *Trier: Center for Research Data in Psychology: PsychData of the Leibniz Institute for Psychology Information ZPID*. DOI: https://doi.org/10.5160/psychdata.zrce16dy99_v20000

**Zygar-Hoffmann, C., Hagemeyer, B., Pusch, S.,** & **Schönbrodt, F. D.** (2020). A large longitudinal study on motivation, behavior and satisfaction in couples: Research data from a four-week experience sampling study with a pre-, post-, and one-year follow-up-assessment. (Version 1.0.0) [data and documentation]. *Trier: Center for Research Data in Psychology PsychData of the Leibniz Institute for Psychology Information ZPID*. DOI: https://doi.org/10.5160/psychdata.zrce18mo99



**Peer review comments**
**The author(s) of this paper chose the Open Review option, and the peer review comments can be downloaded at:**
http://doi.org/10.1525/collabra.278.pr

**How to cite this article:** Zygar-Hoffmann, C., & Schönbrodt, F. D. (2020). Recalling Experiences: Looking at Momentary, Retrospective and Global Assessments of Relationship Satisfaction. *Collabra: Psychology*, 6(1): 7. DOI: https://doi.org/10.1525/collabra.278

**Senior Editor:** Simine Vazire

**Editor:** Simine Vazire

**Submitted:** 02 August 2019 **Accepted:** 01 December 2019 **Published:** 22 January 2020



UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

*Collabra: Psychology* is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS