Evaluating multi-episode events: A boundary condition for the peak-end rule

Talya Miron-Shatz

Woodrow Wilson School for Public and International Affairs, Princeton University

327 Wallace Hall, Princeton University, 08544-1013

Phone: 609 258 7393

Fax: 609 258 5974

Email: tmiron@princeton.edu

Word count: 5,231

Running head: EVALUATING MULTI-EPISODE EVENTS

Evaluating multi-episode events: A boundary condition for the peak-end rule

For Masked Review (revision)

Word count: 5,231

Abstract

This study advances our understanding of the people arrive at the retrospective evaluation of a complex experience. Large samples from the US, France, and Denmark (810, 820 and 805 participants, respectively) reported their sensations during each episode of the previous day using the Day Reconstruction Method (DRM).

The duration-weighted aggregation of these sensations represented the normative approach to evaluation, and, contrary to the predictions of the peak-end rule, it was the best predictor of retrospective evaluations of the day. Two variables represented the heuristic approach, denoting participants' voluntary reports of a wonderful (peak) and awful (low) moment during the previous day. The reported lows, and to a lesser degree, the reported peaks, added to the predictions of the normative variable, especially when predicting comparative rather than absolute evaluations of the previous day. The end episode did not predict retrospective evaluations. The results offer boundary conditions for the peak-end rule, suggesting that retrospective evaluations of complex events rely on the averaged ratings of emotions, ignore ends, and also consider the presence of lows, and occasionally peaks, as subjectively defined by those experiencing them.

Key words: Peak-end rule; judgment; well-being, heuristics, Day Reconstruction Method.

Caroline had a long day: she woke up early, exercised and took a shower. She then got dressed, drove to work, spent over eight hours at the office, did some grocery shopping, cooked, had dinner in front of the television and went to bed. When her best friend calls her the next morning asking 'how was your day' what will Caroline say? Apart from being the naturally occurring unit of our lives, days constitute a relevant example of the complexity of our experiences. During a day we flow in and out of activities, constantly shifting between locations, partners and foci, feeling a wide array of emotions, and it is unclear how we sum all this information into a retrospective evaluation, such as the one Caroline is asked to administer.

Normative and heuristic models have been suggested to answer the question of how people evaluate events. This paper takes on the ambitious task of using both models to unravel what determines the retrospective evaluation of multi-episode real-life events – a description that fits most life experiences – days included. It uses the peak-end rule as a heuristic model for doing so, though taking into account that it may need to be modified in order to be predictive beyond the single-episode context in which it was developed.

Normative models posit that people form moment-by-moment judgments of experiences. The summations or integrals of these momentary judgments lead to holistic judgments. Kahneman, Wakker and Sarin define instant utility as "the pleasure or distress of the moment" (1997, p. 379). Total utility is "constructed from temporal profiles of instant utility according to a set of normative rules" (p. 376). Another determinant of total utility is "the retrospective evaluation of a temporally extended outcome" (p. 379). If the normative assumptions are met, a duration-weighted measurement of how a person felt throughout a multi-episode experience would be the best predictor of his retrospective

evaluation of the experience. Thus, Caroline would multiply the affective value of each of her daily episodes by its relative duration and aggregate the products to evaluate how her day went.

Heuristic, bounded-rationality models (Simon, 1957; Tversky & Kahneman, 1973) challenge the normative stand. They do not assume that computations determine judgments. Rather, people's judgments and evaluations rely on mental shortcuts, and are often based on segments that represent the whole experience, even if the representation is inaccurate. The peak-end rule demonstrates how people use experienced utility to make retrospective heuristic judgments. This descriptive rule states that judgment and recollection of a past event are based on one's sensations when the experience reached extreme intensity, either positive or negative, and when the event ended. These distinct sensations override a duration-weighted average of ongoing affective reports throughout the entire experience, as prescribed by the normative approach (see Bell, Raiffa, & Tversky, 1988, for a comparison of these models). Thus, the heuristic assumption is that when Caroline is asked about her day, she will base her judgment on moments that aroused extreme sensations--such as finally incorporating a handstand in her morning exercise routine, and on the end of the event. In this case, the joy of the handstand and the serenity of going to sleep would lead her to conclude that she had a good day.

A convincing manifestation of the peak-end rule involved colonoscopy, a harrowing medical procedure (Redelmeier & Kahneman, 1996; Redelmeier, Katz, & Kahneman, 2003). In a randomized trial, some patients experienced a typical colonoscopy, while others underwent a modified, slightly longer procedure that ended less painfully. Compared with the regular procedure, the modified procedure was

associated with a significantly lower average of the experienced peak of pain, a less painful memory of the procedure, and an increased likelihood of returning for another colonoscopy. Patients were aware of the event's length, but did not incorporate this into their judgments, which implies duration neglect and reveals the discrepancy between experience and evaluation. Similar findings emerged for other instances of physical discomfort over one event (Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993; Schreiber & Kahneman, 2000) as well as over extended periods of time for post-operational or rheumatic patients (Jensen, Martin, & Cheung, 2005; Stone, Schwartz, Broderick, & Schiffman, 2005, respectively). Additionally, the rule was demonstrated for pleasant stimuli (Fredrickson & Kahneman, 1993). The importance of ends was further manifested by breaking events up at their lowest point, so that the low converged with a local end point. This led to a lower evaluation of the events than when they were broken up at their maximal point (Ariely & Zauberman, 2003).

Examining the peak-end rule in the evaluation of days required three major adaptations to previous investigations. First, whereas in colonoscopy studies, 'peak' denoted 'peak of pain', entire days consist of diverse experiences, evoking a gamut of emotions, so extreme emotionality can be either positive or negative. A person could even report extreme emotionality of both valences, which would not necessarily cancel each other out. Research indicates that it is important to distinguish emotional valence. Isen (1999) adamantly makes the point that positive affect is not the reverse of negative affect. Rather, they differ in their structure and in their impact on cognition, memory and motivation. Positive and negative stimuli even lead to different brain activation (Ito, Larsen, Smith, & Cacioppo, 1998). And while positive events are often rated as more

important, they impact self-esteem less than negative events (Baumeister, Bratlavsky, Finkenauer, & Voss, 2001). Thus, I denoted moments of intensive positive and negative affect as 'peaks' and 'lows,' respectively.

Second, previous studies deduced each participant's peak and low experiences from the affective ratings (for example, of pain). This study was premised on the phenomenological assumption that an experiencing individual is best equipped to determine the peaks and lows in her own life. To this effect, I asked participants to note the presence of peak and low experiences, what these were and when they occurred. The phrasing was overstated, defining peaks as moments that were "unusually wonderful or thrilling", and lows as "unusually unpleasant or awful". Previous studies used mellower phrasings, e.g. "things that stand out to you as either good or bad events that occurred" (Langston, 1994, exp. 2, p. 1120). Yet this also elicited responses that were coded as affectively neutral, so that events of lower intensity obscured the effects of the more prominent moments.

An alternative to directly asking the participants to identify moments of extreme emotionality would be for uninvolved coders sometimes rate the intensity of events (Almeida, Wethington, & Kessler, 2002; Wood, Heimpel, & Michela, 2003). While the uniformity of this method can be an advantage, similar events are not necessarily experienced in a uniform manner. Studies of the military illustrate how commanders can reframe seemingly obvious stressors such as high workload and render them endurable (Britt, Davidson, Bliese, & Castro, 2004). Furthermore, the perceived severity of stressors decreases with education (Almeida, Neupert, Banks, & Serido, 2005) and age has a buffering effect on stressor severity ratings (Almeida & Horn, 2004). This suggests

that intensity ratings could lead to baffling conclusions, so our participants did not rate the intensity of these events. We examined this subjective and explicit approach to peaks and lows by comparing it with a more traditional measurement of peak and low points in the numeric ratings of emotions throughout the episodes.

The third modification to previous investigations of the peak-end rule involved the manner of data collection. Previous studies included a moment-by-moment measurement of affect, which could be conducted over an entire day. Instead, in this study participants used a pen and paper questionnaire, the Day Reconstruction Method (DRM, Kahneman, Krueger, Schkade, Schwarz, & Stone 2004a, 2004b) to record their affect during each episode of the previous day. In the discussion section I elaborate on the methods and its correspondence with experience sampling.

Measuring peaks and lows

In current versions of the DRM (Kahneman, Schkade, Fischler, Krueger, & Krilla, 2007), participants indicated whether they experienced a peak and/or a low, and what rendered these moments so exceptional. Real-life instances of extreme emotionality are hard to manipulate experimentally. A methodological alternative was to inspect the prevalence and impact of peaks and lows in large samples from three countries. Extending the peak-end rule to multi-episode experiences, I hypothesized that the variable marking the presence or absence of peaks and lows would significantly add to the prediction of the previous day's evaluations based on the duration-weighted episode ratings. The variables I used are explained in the Appendix.

Previous studies of the peak-end rule applied a statistical approach to determine when the peak (which was of either positive or negative valence) occurred and the extent

of emotion it was associated with. To emulate this in the present study, I denominated the highest episode affective rating as the *statistical peak*, and the lowest as the *statistical low*. From a phenomenological perspective, the statistical peak would not necessarily correspond with the participant-identified peak, and likewise for the lows. At the core of this investigation were the peaks and lows that the participants singled out.

Ends are very meaningful when measuring discrete events. Consider, for example, aversive episodes like colonoscopy, when the end signifies the termination of pain or discomfort (Redelmeier et al., 2003). Similarly, the "James Dean effect" (Diener, Wirtz, & Oishi, 2001) illustrates how a wonderful life that reached an untimely end received higher ratings than the same life followed by a few moderate years. These examples represent events which constitute coherent wholes, and in which the end is not followed by the immediate subsequent beginning of a similar event. There is little theoretical or practical rationale to assume that the end episode of an entire day would signify such a clear trend (Lowenstein & Prelec, 1993; Varey & Kahneman, 1992; Zauberman, Diehl & Ariely, 2006). Nevertheless, I decided to examine the peak-end rule as a means of evaluating multi-episode experiences, if only to display that a boundary condition exists to the rule, such that the end is immaterial to evaluations when the experiences are not of a coherent nature. Hence, this study explored the contribution of the affective value of the final episode of the day to overall evaluation.

Previous studies demonstrated the peak-end rule using both absolute and relative evaluations (Redelmeier et al., 2003). Similarly, participants made two retrospective evaluations of the previous day--an absolute one calling for a detailed emotional account of the day (e.g., "overall, how happy were you yesterday?"), and an evaluation comparing

it with other days. The comparative evaluation might have more ecological validity as prospect theory suggests that preferences depend on the reference point, rather than remaining invariable (Kahneman & Tversky, 1979).

In summary, this study examined whether multi-episode experiences are evaluated, absolutely and comparatively, using normative or heuristic models, specifically the peak-end rule. The research questions were whether people apply normative or heuristic thinking when evaluating their day, and whether objectively or subjectively defined peaks and lows would be better predictors of the retrospective evaluation. The research hypotheses were that heuristic evaluation will prevail in comparative judgments, and that the affective value of the end episode would not predict the overall evaluation of days.

Method

Participants

The participants were 810 women from Columbus, Ohio, 820 from Rennes, France, and 805 from Odense, Denmark. They were recruited by survey companies, using random-digit dialing. Their mean age was 42.30 in the US (*SD* = 10.94), 38.61 in France (*SD* = 11.23) and 40.26 in Denmark (*SD* = 11.60). The samples were comparable with regards to major demographic variables, such as the percentages of women who are married or cohabiting (69.8%, 60.9%, and 71.9%, respectively), who have a regular job (64.4, 60.8, and 71.9, respectively), or have a biological child at home (55.4%, 52.1%, and 48.6%, respectively). All participants spoke the dominant language at home. They filled out the questionnaires individually.

Materials and Procedure

The participants followed the DRM protocol (Kahneman et al., 2004a, 2004b). They reconstructed the episodes of the previous day, from when they woke up to when they went to sleep. For reasons of feasibility, episodes were specified to last between twenty minutes and two hours. Participants could decide how many episodes to define. For each episode they indicated the location, starting and finishing time, their actions and partners and the extent to which they experienced various feelings, from 0 (*not at all*) to 6 (*very much*). They then rated the previous day overall on the same emotional scales and noted whether the previous day was typical for that day of the week. Finally, they reported whether there was "a moment that was unusually wonderful or thrilling" that day (a peak) and what made it so great, and likewise "unusually awful or unpleasant" (a low) and what made it so bad.

Measures

Measuring the experience

Several variables were derived from the participants' rating of how they felt during each episode measured experience. These variables all ranged from – 6, denoting highly negative affect, to 6 (highly positive affect).

Episode net affect was the average of positive emotions (happy, friendly, calm) minus the average of negative emotions (angry, tense, depressed) as rated for a specific episode.

Duration-weighted net affect measured the participants' reported emotions throughout the day. It was calculated by multiplying the net affect of each episode with the proportion of time it took out of one's waking hours and adding up these products.

Statistical peak and low were the highest and lowest episode net affect ratings for each participant, respectively.

Two variables measured the presence of a peak or a low in the participant's day, as subjectively indicated by her.

Reported Peak was a binary measure, with 1 noting that the person reported having an unusually wonderful or thrilling moment, and 0 noting the absence of such a report.

Reported Low was a binary similar measure: 1 noted that the person reported having an unusually awful or bad moment, and 0 noted the absence of such a report.

Measuring the retrospective evaluation

Two variables measured the retrospective evaluation of one's day.

Yesterday net affect was the average of positive emotions minus the average of negative emotions as rated for the day as a whole (e.g., "overall, how tense were you yesterday?"). This formed the absolute evaluation.

Typical measured how the day compared to what that day of the week usually is, using a scale of 1(*much worse*) to 5 (*much better*). This formed the comparative, relative judgment.

Results

The three samples yielded similar results. However, it is possible that participants in each country used the scales slightly differently, so the samples were analyzed separately.

This inquiry was based on the assumption that people would identify and report instances of intense emotionality in the experiences of their previous day. Reports of

peaks and lows were prevalent despite being optional, which suggests that the concept of peaks and lows is integral to people's experiences. Over half the participants in the US, France and Denmark reported a peak (52.9%, 70.7% and 58%, respectively). A slightly smaller proportion (46.2%, 47.1% and 31%) reported a low.

The method of reporting the peaks and lows required that the participant write when the peak or low happened, and what made it so great (or bad) but not when the peak or low ended. This made the overlap between peak (or low) moments and episodes hard to detect, if such existed. Every effort was made to detect the episode in which the peak or low occurred, using both the time in which they happened, and their content (for example: taking care of a grandchild, or having a family quarrel). Still, one should bear in mind that the peak or low could have taken up a small portion of the episode. Table 1 lists the correlations between all the variables included in the analyses. It indicates that the net affect of the episode in which the reported peak was embedded, and the highest episode net affect (statistical peak) correlate at a level of between .61 (US) and .71 (Denmark), and the correlations for the netaffect of the episode in which the reported low was embedded, and the lowest episode net affect are slightly higher (.72 in the US and Denmark, .75 in France). This suggests a certain convergence of the reported peak and lows with their statistical equivalents, despite the conceptual differences.

- - - - - - - - - - - - - - - - - - -

Enter Table 1 about here

- - - - - - - - - - - - - - - - - - -

Predicting the Retrospective Evaluation of Days

To compare the heuristic peak-end rule with the normative model (represented by the duration-weighted net affect), I included the duration-weighted net affect, the reported peak and low variables, and the net affect of the end episode as predictors of the day's retrospective evaluations. As Table 2 indicates, the experienced duration-weighted net affect was most highly predictive of the retrospective net affect rating of the previous day (β = .84 in the US and Denmark, β = .82 in France, $p < .001$). When considering this, one should take into account that duration-weighted and retrospective netaffect ratings shared the same structure. Furthermore, the participants gave their netaffect retrospective evaluations of the previous day immediately after rating the episodes on the same scales, which could account for some of the shared variance.

- - - - - - - - - - - - - - - - - - -

Enter Table 2 about here

- - - - - - - - - - - - - - - - - - -

Table 2 also illustrates that the 'low' variable consistently added to the predictions of the retrospective evaluations of the previous day, so that the presence of a 'low' was associated with lower evaluations (β = - .10 in the US, β = - .12 in France, and β = - .13 in Denmark, $p < .001$). The presence of peak was only predictive of retrospective evaluations in the US (β = .006, $p < .001$).

The duration-weighted net affect was also the best predictor of the retrospective comparative evaluation ('typical'), though with lower betas than for the overall evaluations (β = .25 in the US, β = .28 in France, and β = .29 in Denmark, $p < .001$). Once again, 'low' was a better predictor than 'peak', except in the US.

The end episode did not add to the participants' overall evaluations of the previous day beyond its contribution to the duration-weighted net affect. It was only a significant predictor of the comparative ('typical') evaluations in the US (β = - .08, $p <$ .05).

Previous studies of the peak-end rule used statistically determined peaks and lows. Likewise, Table 3 displays an analysis of the extent to which the net affect of the statistical peak and low episodes added to the prediction of the day's evaluations.

- - - - - - - - - - - - - - - - - - -

Enter Table 3 about here

- - - - - - - - - - - - - - - - - - -

As indicated by Table 3, the statistical peak and especially the statistical low were good sole predictors of *typical* and excellent predictors of yesterday net affect. Once duration-weighted net affect was entered in the regression as a predictor, however, the statistical peak only added to the prediction of 'typical' for the French sample, though the beta value was negative. The statistical low added significantly to the predictions of yesterday net affect. However, it did not add to the prediction of 'typical' in France. Thus, adding the presence of reported peaks and lows to duration-weighted net affect results in stronger predictions of previous day evaluations than adding the net affect values of the statistical peak and low episodes.

The Nature of Reported Peaks and Lows

Reported peaks and lows are pivotal in this investigation, so a brief discussion of their nature is warranted. These instances seldom revolved around extraordinary events,

and yet were meaningful enough for the participants to single them out and report them, when they could have moved on to the next question.

When relating what made a peak exceptional, the first responses in the US (sorted by participant number) were "rocking the baby [grandchild]", and "calculations were right for the first time." French peaks started with "returning home to my husband", and "shopping". The Danes mentioned "two deer passed by us," and "relaxing in the yard, doing what I like to do."

Similarly, when recounting what made the lows so bad, American participants reported: "my husband and I have been fighting a lot and I was wondering if our marriage will last", and "computer crashed." Some French lows were: "had to pay bills", and "fast food dinner even though I was on a diet." The Danes mentioned "allergic reaction", and "there was a mess everywhere and it was dirty." This gives a taste of what people identified as their peaks and lows, which merit a more systematic investigation.

Discussion

This paper examined whether the retrospective evaluation of multi-episode events, focusing on days, the naturally occurring unit of our lives, was normative or heuristic. Specifically, I used a normative approach by aggregating and duration-weighting the sensations associated with the previous day's episodes, and a heuristic approach where the participants identified instances of extreme positive and negative emotionality during their day (peaks and lows). Contradictory to the predictions of the peak-end rule, the normative variable was the best predictor of retrospective evaluations of the previous day, both absolute and comparative. The presence of lows, and to a lesser degree, of peaks, added to the retrospective evaluations of the previous day that were

generated using the normative model. The normative variable, duration-weighted net affect, however, was composed of about a 100 observations per participants, whereas the presence of peaks and lows was marked by two binary variables. The affective value of the end episodes rarely added to the prediction of retrospective evaluations, perhaps because the end of the day does not carry a distinct emotional meaning. Large samples from three countries yielded similar results, though future investigation could shed light on the intricacies of retrospective evaluations in each country. The results suggest a normative evaluation of multi-episode events, supported by a heuristics reliance on lows and peaks, but not ends. This is an important contribution to the understanding of how complex events are evaluated in retrospect.

The main criticism that arises from the study is that the entire measurement – of experience and retrospective evaluation – was done after the fact. However, DRM instructions encouraged participants to take the time to re-live each episode in detail-- their activities, who they were with, and their own feelings. This evokes the contextual experience, as opposed to the semantic and decontextualized remembering self which involves one's beliefs about emotions (Robinson & Clore, 2002). The DRM replicated affective patterns obtained with experience sampling, in which participants are asked their feelings as they are experiencing them during random moments throughout the day (Stone, Shiffman, & DeVries, 1999). In particular, both forms of data collection were associated with a V shaped diurnal rhythm for fatigue. This attests to the validity of the DRM findings, as it diverged from the lay conviction that fatigue increases steadily throughout the day (Stone et al., 2006). The above evidence allows one to consider the affective ratings of episode as experiential, even though they were done the following

day. An additional caveat of this particular journal method was that the episodes had to last at least 20 minutes, so briefer instances would be lost. The form of reporting peak and low moments, however, served as a partial compensation, as it allowed for noting events that were subjectively meaningful to the participant, regardless of their length.

The proportion of the contribution of heuristic and normative variables depended on the type of retrospective prediction. Absolute evaluations, such as estimating how friendly, angry, or depressed you were yesterday call for a thorough scrutiny of the previous day and elicit the experiencing self, especially when retrospective and episode-based evaluations share the same structure. Therefore, these evaluations relied heavily on the normatively-generated duration-weighted net affect variable. Deciding whether yesterday was better or worse than a typical day emphasizes the perspective of the remembering self. This form of evaluation still relied on the normative evaluation, but relied to an almost similar extent on the heuristic notion of the lows in ones day, and to a lesser extent – on peaks.

Comparative evaluations, such as the question of how yesterday compared to a typical day, are pivotal to judgments of abilities and attributes, well-being, self-esteem and selection, among other factors (Miller & Prentice, 1996). Recent examples show comparative judgments in self-construals, body image, and consumer behavior (Guimond, Chatad, Martinot, Crisp, & Redersdorff, 2006; Shroff & Thompson, 2006;, and Karlsson, Garling, Dellgran, & Klingander, 2005, respectively). Comparisons also carry important emotional outcomes: the prospective pleasure and pain conceived by comparing various plausible results affects decisions and choices (Mellers, 2000). Thus, standards are actively constructed through comparison, rather than remaining constant

regardless of the context. Parkinson, Briner, Reynolds, and Totterdell (1995) also found that peaks and lows contributed significantly to comparative evaluations of typicality. They attributed this to the fact that peaks and lows call for unusual events, which may also be atypical. Yet participants in the present study rated the weekend days significantly higher on the 'typical' scale than weekdays, hinting that typicality also provides a general evaluation of the day.

Former studies of the peak-end rule were criticized as confronting two key limitations: using "affect-inducing stimuli… [that are] fairly uniform, likely to produce variations in valence and intensity, but not in specific emotions" and measuring continuous real-time ratings on one dimension, rather than "multiple discrete emotions" (Fredrickson, 2000, p. 594). The present study overcame the first limitation by examining entire days, which are anything but uniform. Open-ended reports of peaks conveyed a rich emotional array, including tranquility, achievement, love, pleasure, friendship, and faith. Likewise, lows conveyed loneliness, hostility, frustration, anger, fatigue and anxiety. The second limitation was overcome by allowing the participants to rate ten emotions for each episode, six of which were compiled into the net affect variable.

Another limitation of past investigations was their restricted context. Current advocacy calls for examining "personality in its natural habitat" and listening to people's daily experiences (Mehl, Gosling, & Pennebaker, 2006, p. 862), as well as letting people use their own words to delineate the troughs and summits in their life (Diener et al., 2001). Consistent with this, the present exploration was based on people's subjective definition of peaks and lows. The participants were at liberty to decide which elements were "unusually wonderful or thrilling" or "unusually awful or unpleasant," if any. Most

of the peaks and lows revolved around events that may seem mundane, but were substantial enough to be singled out by the participants, as moments of unique emotionality. Consider, for example, "my husband won 2 trees in a silent auction" (a peak), and "going to Wal-Mart" (a low). Furthermore, allowing the participants to highlight such instances of varying length helps alleviates a major caveat of the DRM, namely that episodes should last at last twenty minutes, so that events that are meaningful but brief, are not registered.

The end is one of an event's defining features (Ariely and Carmon, 2003), as are the rate and valence of changes and moments of extreme intensity. Most human experiences are goal-directed, which explains why ends should be given extra weight (Spiegel, 1998). In many cases, such as queuing (Carmon & Kahneman, 1996), the end dominates the situation: there are no benefits for partial completion, and unless the end is successful, the experience needs to be repeated. This is not necessarily true for an entire day, however. The days which the participants in this study evaluated do not resemble this scenario: the end does not determine the outcome of the situation, if a day can be said to have an outcome at all, and the end is not finite in the sense that more days are to follow. Therefore, and contrary to the predictions of the peak-end rule, the affective value of the participants' final episode added null predictive value to the retrospective evaluations of the previous day. Yet in the past, the peak-end rule was mostly demonstrated for events which displayed a clear trend. Partitioning events so they displayed an ascending or descending trend increased the weight of the end in overall evaluations (Ariely & Zauberman, 2003). This suggests that another partitioning that did not entail a clear trend would result in a reduced weight for the end. Going back to

Caroline, her final episode involved going to sleep, which carried scarce input for the evaluation of the day. This is to suggest that the end component of the peak-end rule has boundary conditions which have not been made explicit in previous studies.

A key corollary of the peak-end rule is duration neglect, an insensitivity to the duration of the event, as demonstrated by basing the rating on two time points and ignoring the rest. Our participants displayed this tendency, as the correlation between the duration-weighted average of their affective episode ratings (Netaffect) was almost identical to the non-duration-weighted average. This could imply that arriving at the episode rating involves a local process of relying on segments of the event (in this case, the episode), rather than using a moment by moment evaluation.

There could be a few reasons for the disconfirming of the peak-end rule hypothesis in this study, as suggested by the superior ability of a normative variable to predict retrospective evaluations of the previous day, compared with the predictive ability of peaks and lows, and with the null predictive ability of ends. Stone et al. (2005, p. 1340) named the parsimonious tendency to base judgment on salient aspects or of one's experience the "peak (or salience) memory heuristic". However, in the present study, participants reconstructed every episode of their previous day, so memory could not have prevented them from basing their judgment on multiple pieces of information, leading to a more normative assessment. Within the heuristic framework, affect-driven mechanisms of attention determine which features of stimuli are given weight in retrospective evaluation (Finucane, Peters, & Slovic, 2003). Peaks and lows are natural candidates. Ends, on the other hand, are not very helpful when studying complex events which do not display cohesive trends, so that their end is devoid of a defining meaning. The increased

reliance on the presence of lows, compared with peaks, is consistent with literature regarding the greater effect of negative events and procedures relative to positive ones (Baumeister et al., 2001; Kanouse, 1984; Van den Bos, K., & Van Prooijen, J.W., 2001).

Discrepancies between experience and evaluation often result in a failure to choose or predict options that will maximize experienced happiness (Hsee & Hastie, 2006; Wirtz, Kruger, Napa Scollon, & Diener, 2003). The findings suggest that the gap between experience, as captured by the episodes that make up a day, and memory, as captured by how that day is evaluated, is in fact quite small. This is particularly true for overall, rather than comparative evaluations. The findings further suggest that the memory-experience gap and duration neglect that underlined the peak-end rule, may be less characteristic of the evaluation of multi-episode events.

References

Almeida, D. M., Horn, M. C. (2004). Is Daily Life More Stressful during Middle Adulthood? In O. G. Brim, C. D. Ryff, & R. C. Kessler (Eds.), *How healthy are we?: A national study of well-being at midlife* (pp. 425-451). Chicago: University of Chicago Press.

Almeida, D. M., Neupert, S. D., Banks, S. R., Serido, J. (2005). Do daily stress processes account for socioeconomic health disparities? *Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 60B*(2, SpecIssue), 34-39.

Almeida, D. M., Wethington, E., Kessler, R. (2002). The Daily Inventory of Stressful Events: An interview-based approach for measuring daily stressors. *Assessment*, *9*(1), 41-55.

Ariely, D., & Carmon, Z. (2003). Summary assessment of experiences: The whole is different from the sum of its parts. In G. Lowenstein, D. Read, & R. Baumeister (Eds.), *Time and decision: Economic and psychological perspectives on intertemporal choice* (pp. 323-349). New York, Russell Sage Foundation.

Ariely, D., & Zauberman, G. (2003). Differential partitioning of extended experiences. *Organizational Behavior and Human Decision Processes, 91*(2), 128-139.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001) Bad is stronger than good. *Review of General Psychology, 5*(4), 323-370.

Bell, D. E., Raiffa, H., & Tversky, A. (1988). Descriptive, normative, and prescriptive interactions in decision-making. In D. Bell, H. Raiffa, & A. Tversky (Eds.), *Decision Making* (pp. 9-30). New York: Cambridge University Press.

Britt, T. W., Davison, J., Bliese, P. D., Castro. C. A. (2004). How Leaders Can Influence

the Impact That Stressors Have on Soldiers. *Military Medicine*, *169*(7), 541-545.

Carmon, Z., & Kahneman, D. (1996). The experienced utility of queuing: real time affect

and retrospective evaluations of simulated queuing. Working paper, Duke

University.

Diener, E., Wirtz, D., & Oishi, S. (2001). End effects of rated quality of life: The James

Dean effect. *Psychological Science, 12*(2), 124-128.

Finucane, M. L., Peters, E., & Slovic, P. (2003). Judgment and decision making: The

dance of affect and reason. In S. L. Schneider, & J. Shanteau (Eds.), *Emerging*

perspectives on decision research (pp. 327-364). New York: Cambridge

University Press.

Fredricksen, B. L. (2000). Extracting meaning from past affective experiences: The

importance of peaks, ends, and specific emotions. *Cognition and Emotion, 14*(4),

577-606.

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations

of affective episodes. *Journal of Personality and Social Psychology, 65*(1), 45-55.

Guimond, S., Chatad, A., Martinot, D., Crisp, R. J., & Redersdorff, S. (2006). Social

comparison, self-stereotyping and gender differences in self-construals. *Journal of*

Personality and social Psychology, 90(2), 221-242.

Hsee, C. K., & Hastie, R. (2006). Decision and experience: Why don't we choose what

makes us happy? *Trends in Cognitive Sciences, 10*(1), 31-37.

Isen, A. M. (1999). Positive affect. In T. Dalgleish & M. J. Power (Eds*.), Handbook of*

cognition and emotion (pp. 521-539). New York: John Wiley & Sons Ltd.

Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information

weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*, *75*(4), 887-900.

Jensen, M. P., Martin, S. A., & Cheung, R. (2005). The meaning of pain relief in a clinical trial. *Journal of Pain, 6*(6): 400-406.

Kahneman, D. Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4*(6), 401-405.

Kahneman, D., Krueger, A.B., Schkade, D., Schwarz, N., & Stone, A. A. (2004a). A survey method for characterizing daily life experiences: The Day Reconstruction Method. *Science, 306*(5702), 1776-1780.

Kahneman, D., Krueger, A. B., Schkade, D., Schwarz, N., & Stone, A. A. (2004b). The Day Reconstruction Method (DRM): Instrument documentation. Retrieved April 3, 2005, from Science Magazine Web site: http://www.sciencemag.org/cgi/content/full/306/5702/1776/DC1.

Kahneman, D., Schkade, D. A.., Fischler, C., Krueger, A. B., & Krilla, A.C. (2006). The structure of well-being in two cities. Working paper, Princeton University.

Kahneman, D., & Tversky, A. (1979). Prospect theory. *Econometrica, 47*, 263-292.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics, 112*(2), 375-405.

Kanouse, D.E. (1984). Explaining negativity bias in evaluative and choice behavior: Theory and research. In: Kinnear, T. (Ed.). *Advances in Consumer Research*, Vol. 11. Ptovo: UT: Association for Consumer Research. pp. 703-308.

Karlsson, N., Garling, T., Dellgran, P., & Klingander, B. (2005). Social comparison and

consumer behavior: When feeling richer or poorer than others is more important than being so. *Journal of Applied Social Psychology, 35*(6), 1206-1222.

Langston, C. A. (1994). Capitalizing on and coping with daily-life events: Expressive responses to positive events. *Journal of Personality and Social Psychology, 67*(6), 1112-1125.

Loewenstein, G., & Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review, 100*, 91-108.

Mehl, M. R., Gosling, S. D., & Penebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90*(5), 862-877.

Mellers, B. A. (2000). Choice and the relative pleasure of consequences. *Psychological Bulletin, 126*(6), 910-924.

Miller, D. T., & Prentice, D. A. (1996). The construction of social norms and standards. In E.T. Higgins & A.W. Kruglansky (Eds.), *Social psychology: Handbook of basic principles* (pp. 799-829). New York: Guilford Press.

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.

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*(1), 3-8.

Redelmeier, D. A., Katz, J., & Kahneman, D. (2003). Memories of colonoscopy: A randomized trial. *Pain, 104*(1-2), 187-194.

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an

accessibility model of emotional self-report. *Psychological Bulletin, 128*, 934-

960.

Schreiber, C. A., & Kahneman, D. (2000). Determinants of the remembered utility of

aversive sounds. *Journal of Experimental Psychology: General, 129*(1), 27-42.

Seligman, M. E. P., Steen, T. A., Park, N., & Peterson, C. (2005). Positive psychology:

Positive psychology progress: Empirical validation of interventions. *American

Psychologist, 60*(5), 410-421.

Shroff, H., & Thompson, J. K. (2006). The tripartite influence model of body image and

eating disturbance: A replication with adolescent girls. *Body Image*, 3(1), *17-23*.

Simon, H. A. (1957). *Models of Man*. New York: Wiley.

Speigel, D. (1998). Getting there is half the fun: Relating happiness to health.

Psychological Inquiry, 9, 66-68.

Stone, A. A., Schwartz, J. E., Broderick, J. E., & Shiffman, S. S. (2005). Variability of

momentary pain predicts recall of weekly pain: A consequence of the peak (or

salience) memory heuristic. *Personality and Social Psychology Bulletin, 31*(10),

1340-1346.

Stone, A. A., Schwartz, J. E., Schwartz, N., Schkade, D., Krueger, A., & Kahneman, D.

(2006). A population approach to the study of emotion: Diurnal rhythms of a

working day examined with the day reconstruction method. *Emotion, 6*(1):139-

149.

Stone, A. A., Shiffman, S. S., & DeVries, M. W. (1999). Ecological momentary

assessment. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The*

foundations of hedonic psychology (pp.61-84). New York: Russel-Sage Institute.

Tversky, A., & Kahneman, D. (1973). *Judgment under uncertainty: Heuristics and Biases*. Oxford, England: Oregon Research Institute.

Varey, C. A., & Kahneman, D. (1992). Experiences extended across time: Evaluation of moments and episodes. *Journal of Behavioral Decision Making, 5*(3), 169-185.

Van den Bos, K., & Van Prooijen, J.W. (2001). Referent cognitions theory: The psychology of voice depends on closeness of reference points, *Journal of Personality and Social Psychology* **81** (2001), pp. 616–626.

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

Wood, J. V., Heimpel, S. A., Michela, J. L. (2003). Savoring Versus Dampening: Self-Esteem Differences in Regulating Positive Affect. *Journal of Personality and Social Psychology*, *85*(3), 566-580.

Zauberman, G., Diehl, K., & Ariely, D. (2006). Hedonic versus informational evaluations: Task dependent preference for sequences of outcomes. *Journal of Behavioral Decision Making, 19*(3), 191-211.

Table 1

Correlations between major variables

	1	2	3	4	5	6	7	8	9	10
1. Typical										
U.S.[a]	–									
France[b]	–									
Denmark[c]	–									
2. ODNA										
U.S.	.43	–								
France	.43	–								
Denmark	.42	–								
3. DWNA										
U.S.	.34	.87	–							
France	.33	.87	–							
Denmark	.35	.87	–							
4. Peak										
U.S.	.27	.15	.13	–						
France	.13	.14	.15	–						
Denmark	.19	.10**	.12**	–						
5. Low										
U.S.	−.26	−.38	−.34	.11**	–					
France	−.23	−.36	−.30	.17	–					
Denmark	−.30	−.38	−.31	.12	–					
6. EENA										
U.S.	.23	.60	.69	.07*	−.28	–				
France	.20	.58	.64	.09**	−.22	–				
Denmark	.13	.45	.54		−.18	–				
7. PENA										
U.S.	.17**	.44	.56	.[d]	−.12*	.39	–			
France	.14**	.47	.55	.[d]	−.17	.39	–			
Denmark	.17	.53	.63	.[d]	−.11*	.27	–			
8. LENA										
U.S.	.30	.58	.58	.12**	.[d]	.38	.16*	–		
France	.23	.52	.55		.[d]	.36	.20**	–		
Denmark		.60	.63		.[d]	.35	.29	–		
9. SPNA										
U.S.	.21	.59	.72	.15	−.15	.58	.61	.25	–	
France	.14	.60	.72	.25		.51	.64	.23	–	
Denmark	.17	.66	.75	.16	−.08**	.48	.71	.40	–	
10. SLNA										
U.S.	.34	.71	.74		−.43	.48	.31	.72	.31	–
France	.27	.69	.72		−.45	.45	.31	.75	.28	–
Denmark	.32	.70	.76		−.45	.42	.33	.77	.39	–

Note. All ps are < .001 unless marked otherwise. Non-significant correlations are not shown. Typical = yesterday typicality; ODNA = overall day net affect; DWNA =

duration-weighted experienced net affect; Peak = reported a peak (dummy-coded); Low = reported a low (dummy-coded); EENA = end episode net affect; PENA = peak episode net affect; LENA = low episode net affect; SPNA = statistical peak net affect; SLNA = statistical low net affect.

[a]n = 810. [b]n = 819. [c]n = 810. [d]No correlation could be computed because all cases with valid values for the continuous measure had the same value for the dummy variable.

*p < .05. **p < .01.

Table 2

Summary of Regression Analysis for Predicting Retrospective Evaluations of the

Previous Day with Duration-Weighted (DW) Net Affect, Net Affect for the End Episode

and Peak-Low

	United States		France		Denmark	
	B	β	B	β	B	β
Predicted Variable - Overall day net affect (absolute judgment)						
DW Experienced Net affect	1.11	.84***	1.13	.82***	1.05	.84***
	(.03)		(.03)		(.03)	
End Episode Net affect	-.01	-.01	.02	.02	-.02	-.02
	(.02)		(.02)		(.02)	
Reported a Peak	.26	.06***	.15	.03	.06	.02
	(.08)		(.08)		(.06)	
Reported a Low	-.45	-.10***	-.49	-.12***	-.51	-.13***
	(.08)		(.08)		(.07)	
Predicted Variable – Yesterday Typicality (comparative judgment)						
DW Experienced Net affect	.13	.25***	.16	.28***	.21	.29***
	(.02)		(.03)		(.03)	
End Episode Net affect	-.01	-.02	-.01	-.03	-.04	-.08*
	(.02)		(.02)		(.02)	
Peak Episode Net affect	.48	.26***	.23	.12***	.38	.19***
	(.06)		(.07)		(.07)	
Low Episode Net affect	-.39	-.21***	-.31	-.18***	-.54	-.25***
	(.06)		(.06)		(.07	

Note. Standard errors for unstandardized betas are in parentheses.

Table 3

Summary of Regression Analysis for Predicting Retrospective Evaluations of the

Previous Day with Duration-Weighted (DW) Net Affect, Net Affect for the End Episode

and Net Affect for the Statistical Peak and Low Episode

	United States		France		Denmark	
	B	β	B	β	B	β
Predicted Variable - Overall day net affect						
DW Experienced Net affect	1.02	.77***	1.06	.76***	0.95	.75***
	(.05)		(.06)		(.05)	
End Episode Net affect	0.01	.01	.04	.04	-.02	-.02
	(.02)		(.02)		(.02)	
Statistical Peak Net affect	-0.03	-.01	0.02	.01	0.11	.07
	(.05)		(.05)		(.05)	
Statistical Low Net affect	0.12	.14***	0.10	.11***	0.09	.10**
	(.02)		(.03)		(.02)	
Predicted Variable – Yesterday Typicality						
DW Experienced Net affect	0.10	.18*	0.31	.53***	0.32	.45***
	(.04)		(.05)		(.06)	
End Episode Net affect	0.00	-.01	0.00	.00	-0.03	-.07
	(.02)		(.02)		(.02)	
Statistical Peak Net affect	0.02	.02	-0.17	-.24***	-0.14	-.16**
	(.04)		(.04)		(.05)	
Statistical Low Net affect	0.07	.20***	-0.02	-.05	0.03	.07
	(-.02)		(.02)		(.03)	

Note. Standard errors for unstandardized betas are in parentheses.

Appendix: Explaining the variables using Caroline's day

This paper involves several conceptualizations of the peak (and low) notion. To clarify the nature of the variables, I will demonstrate them using the hypothetical Caroline's day. This was the first time Caroline managed to do a hand stand as part of her exercise at home before dinner. The handstand occurred at 6:22 pm and the whole exercise episode took place between 6:10 and 6:30 pm.

Caroline gave the exercise episode ratings which resulted in a Netaffect score of 3.66 (the specific ratings were: calm = 6, happy = 5 friendly = 4 for the positive emotions, and angry = 2, tense = 1, depressed = 1 for the negative emotions).

For the question 'was there a moment yesterday that was unusually wonderful or thrilling' (peak) Caroline noted the handstand as her reported peak. She did not have to rate it, and its presence was marked by a binary variable.

For the question 'was there a moment yesterday that was unusually bad or awful' (low) Caroline did not note anything. She did not have to rate it, and its presence was marked by a binary variable.

Hence, the 'peak episode' was the exercise episode, because the reported peak was temporally embedded in it. Its Netaffect score was 3.66.

Since Caroline did not report a low, she could not have a low episode, and this was a missing value.

For 'statistical peak' I would use the highest Netaffect value of all episode ratings, which was 4.00, for the dinner episode. Note that the Netaffect rating for this episode was higher than that for the episode in which the reported peak was embedded.

Similarly, for 'statistical low' I would use the lowest Netaffect value of all episode ratings, which was 2.00, for one of Caroline's work episodes.

The 'end' would simply be the Netaffect value for Caroline's last episode during the day, when she lay in bed with a book. It received a Netaffect rating of 3.

In the duration weighted (dw) Netaffect of Caroline's day, the Netaffect of each episode would be multiplied by the episode's relative length in Caroline's waking hours. She was up at 7 am and went to bed at 11 pm, resulting in 14 waking hours (or 840 minutes), so, for example, the Netaffect of the exercise episode, lasting 20 minutes, would be divided by 42 when entered in the dw score.