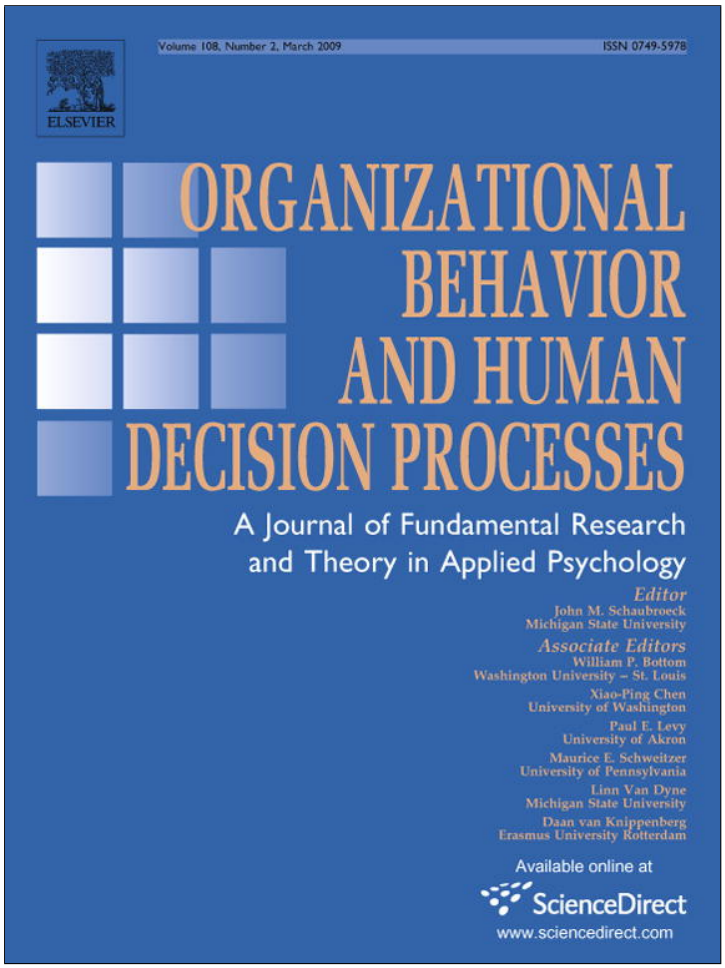


Organizational Behavior and Human Decision Processes 108 (2009) 303–314

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

Duration neglect by numbers—And its elimination by graphs ☆

Michael J. Liersch [a,*], Craig R.M. McKenzie [b]

[a] Leonard N. Stern School of Business, New York University, 40 West 4th Street, 701C, New York, NY 10012, USA
[b] Rady School of Management and Department of Psychology, University of California, San Diego, 9500 Gilman Drive, MC 0553, La Jolla, CA 92093-0533, USA

ARTICLE INFO

Article history:
Received 9 October 2007
Accepted 5 July 2008
Available online 8 November 2008

Accepted by John Schaubroeck

Keywords:
Duration neglect
Peak/end rule
Hedonic experience
Heuristics
Biases

ABSTRACT

People tend to neglect duration when retrospectively evaluating aversive experiences, causing memories to be at odds with experienced pain. However, memory was not involved in the original demonstration of duration neglect. Instead, people evaluated others' experiences represented by lists of discomfort ratings. Duration was said to be neglected because attention was focused on peak and end ratings. Three experiments are reported demonstrating that graphs rather than number lists can make duration neglect disappear without increasing attention to episode duration. Graphs can eliminate duration neglect because, relative to number lists, strategies that incorporate duration are more easily employed. The results suggest that when hedonic information does not have to be remembered, people will use all, not just peak and end, moments when evaluating experiences, and that format presentation affects how people combine those moments. Caution is recommended when making theoretical and prescriptive generalizations based on duration neglect.

After undergoing experiences such as painful medical procedures or unpleasant movies, duration of the event would seem to play an important role in evaluating overall levels of experienced pain. But, surprisingly, when providing impressions of aversive hedonic experiences, people tend to neglect duration (Fredrickson & Kahneman, 1993; Kahneman, 2003; Kahneman, 2000a, 2000b; Kahneman, 1999; Kahneman & Frederick, 2002; Kahneman, Frederickson, Schreiber, & Redelmeier, 1993; Kahneman, Wakker, & Sarin, 1997; Redelmeier & Kahneman, 1996; Redelmeier, Katz, & Kahneman, 2003; Schreiber & Kahneman, 2000; Varey & Kahneman, 1992).

The initial demonstration of this phenomenon had participants evaluate hypothetical unpleasant episodes of differing durations (Varey & Kahneman, 1992). Each aversive episode was presented as a numerical list of discomfort ratings reported by hypothetical subjects in 5-min intervals (see Fig. 1a). Higher ratings represented more discomfort on a scale of 0 to 10 (e.g., an episode rated {2,5,8} depicted a 15 min experience of increasing discomfort). Varey and Kahneman found that duration of the experience had little effect on participants' overall judgments. For example, the episode {2,5,8,4} was rated more favorably than {2,5,8} even though the only difference between the two episodes is that the former has an additional 5 min of pain. The potential implications of such preferences are far-reaching. For instance, when making a prospective choice between two unpleasant medical treatments—e.g., two different chemotherapy regimens—Varey and Kahneman's data suggest that people could choose the treatment that is characterized as longer, and with more total pain.

Why did participants judge a longer, more painful episode to be better than a shorter, less painful one? Kahneman and colleagues have proposed that people *average* peak and end hedonic moments of an experience instead of temporally integrating, or *adding*, all hedonic moments when forming overall evaluations (e.g., Schreiber & Kahneman, 2000). For example, in Varey and Kahneman's (1992) rating task, an average of the peak and end ratings would result in a lower (better) evaluation for {2,5,8,4} relative to {2,5,8} (6 vs. 8), whereas adding ratings at each time interval would reverse this non-normative pattern of judgments (19 vs. 15; for a normative account of *temporal monotonicity*, whereby adding aversive moments should make the experience worse, see Kahneman et al., 1997). Participants are said to attend only to peak and end moments, and not duration, because the worst and last moments are prototypical exemplars of an experience and are therefore most salient (Fredrickson & Kahneman, 1993; Kahneman, 2003; Kahneman, 2000b; Kahneman & Frederick, 2002; Kahneman et al., 1993; see also Ariely & Carmon, 2000; Frederickson, 2000).

Duration-free, peak-ended evaluation patterns are not isolated to situations involving others' hypothetical experiences, but have also been demonstrated in people's retrospective evaluations of their own experiences and when people make choices based on memories of their experiences (Fredrickson & Kahneman, 1993;

☆ This research was supported by National Science Foundation Grant SES-0551225. Some of the results were presented at the 2005 Annual Meeting of the Society for Judgment and Decision Making in Toronto, ON. Rachel Rosenthal and Shlomi Sher provided valuable comments on earlier drafts.

* Corresponding author. Fax: +1 858 534 7190.
E-mail address: mliersch@stern.nyu.edu (M.J. Liersch).

0749-5978/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.obhdp.2008.07.001

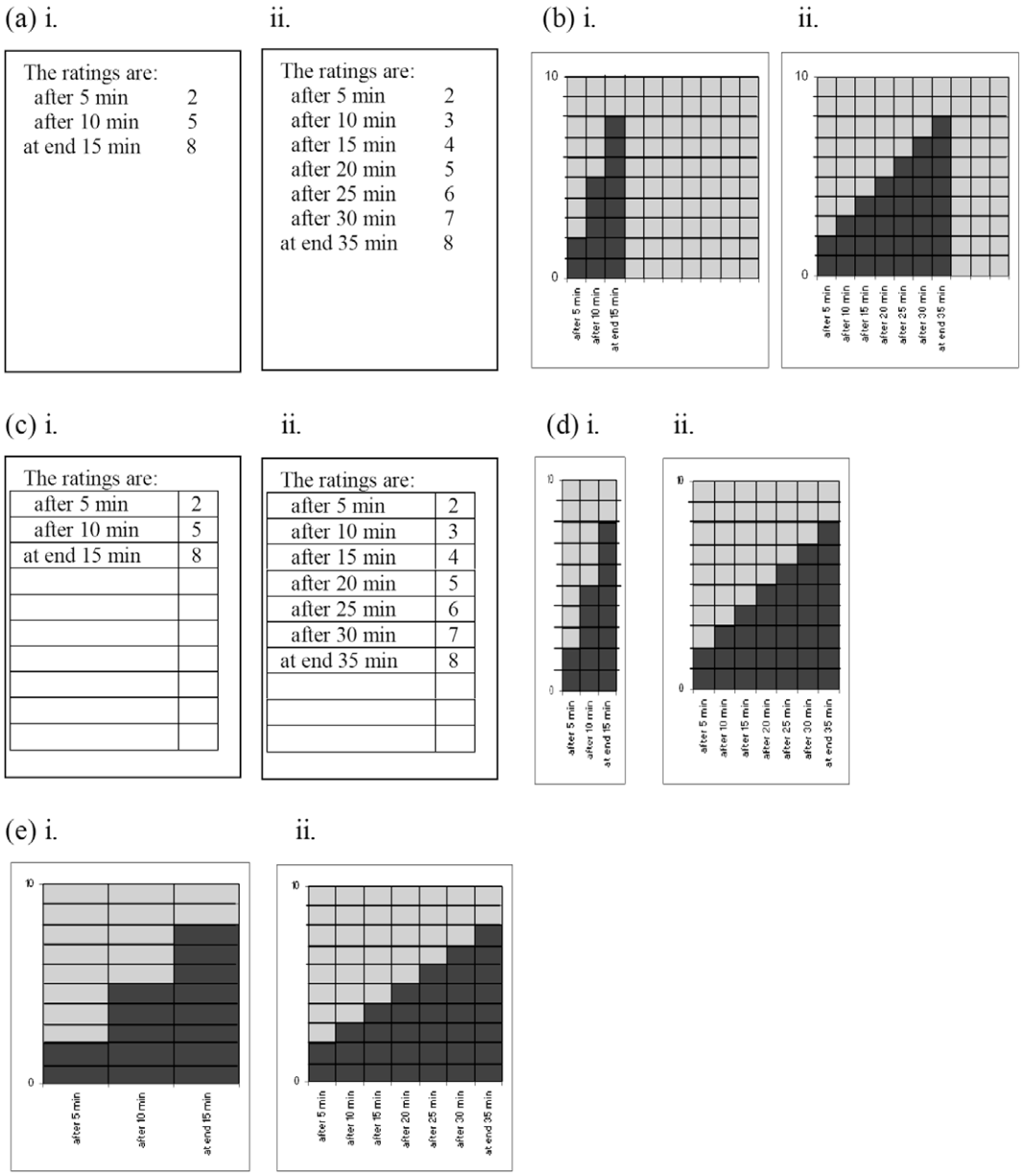


Fig. 1. Sample discomfort sequences {2,5,8} (15 min) and {2,3,4,5,6,7,8} (35 min) in numerical (a), full graphical (b), and modified numerical (c) formats from Experiment 1 and the modified graphical (d) and modified-fixed-axis graphical (e) formats from Experiment 2.

Kahneman et al., 1993; Redelmeier & Kahneman, 1996; Redelmeier et al., 2003; Schreiber & Kahneman, 2000). According to Kahneman et al. (1993), "[P]eople prefer to repeat the experiences that have left them with the most favorable memories—not necessarily the experiences that actually gave the most pleasure and the least pain" (p. 404). As a result, researchers have proposed prescriptions intended to enhance well-being, including lengthening medical procedures by adding a period of diminishing pain (Kahneman, 2000b; Kahneman et al., 1993; Redelmeier & Kahneman, 1996). Indeed, Redelmeier et al. (2003) found that adding a period of diminishing pain to patients' colonoscopies improved retrospective evaluations of those procedures and increased the likelihood that patients would undergo future colonoscopies.

Presumably, people form duration-free, peak-ended evaluations of real experiences because it is difficult to remember and then integrate experienced utility at each moment in time (Fredrickson & Kahneman, 1993; Kahneman et al., 1993; Redelmeier & Kahneman, 1996; Redelmeier et al., 2003; Schreiber & Kahneman, 2000; see also Ariely, Kahneman, & Loewenstein, 2000). For example, if asked how much one liked an unpleasant film, basing the judgment on the worst and last moments seems sensible relative to reconstructing the entire film in memory (Fredrickson & Kahneman, 1993). Importantly, though, in Varey and Kahneman's (1992) rating task of others' hypothetical experiences, *nothing had to be remembered*, so participants could have easily used all hedonic ratings—not just the peak and end.

Furthermore, while attending to only the peak and end moments of an experience seems plausible when duration comparisons cannot easily be made (e.g., colonoscopies), it seems less likely when experiences of differing duration are easily comparable (e.g., evaluating multiple hypothetical experiences of differing durations as in Varey & Kahneman, 1992). Kahneman and Frederick (2002) stated that, while between-participants designs will almost *always* result in duration neglect (where explicit duration comparisons cannot be made), within-participants factorial designs will almost *never* show the effect because attention is brought toward duration (see also Kahneman, 2003; Schreiber & Kahneman, 2000; Ariely et al., 2000; Ariely & Loewenstein, 2000;

Morewedge, Kassam, Hsee, & Caruso, in press). This view predicts that duration *would not* have been neglected in Varey and Kahneman (1992) because participants could compare 24 hypothetical experiences ranging from 15 to 35 min in a within-participants factorial design (for an example, compare Fig. 1a, i. and ii.)—yet duration accounted for only 3% of the variance in evaluations.

In short, then, the factors typically used to explain duration neglect do not appear to account for Varey and Kahneman's (1992) original demonstration of the phenomenon. In this article, we argue that information display (i.e., format) contributed to duration neglect in Varey and Kahneman's hypothetical rating task. Specifically, we hypothesize that the number list format used by Varey and Kahneman led participants to average the numbers, but that graphical formats would encourage adding strategies, thereby eliminating or reducing duration neglect.

Previous research has shown that how information is presented influences how people utilize that information (e.g., Kleimuntz & Schkade, 1993; see also Schkade & Kleimuntz, 1994). For example, in a classic study by Russo (1977), supermarket customers saved about 2% on purchases when unit price information was provided in a list, rather than on separate unit price tags under each product. Russo argued that while unit price information is available to shoppers when presented on separate tags, listing unit prices in one place makes price comparisons relatively easier. Russo and Leclerc (1991) found support for this "ease" argument by showing that unit price lists massively reduced comparison time relative to when unit prices were displayed on separate tags.

More generally, there is a tendency for ease—specifically, people's desire to minimize cognitive effort—to dictate how information is integrated (Payne, Bettman, & Johnson, 1993; see also Chu & Spires, 2003). This tendency is critical to our hypothesis because numerical displays have been shown to differentially affect the ease of information use compared to other information displays, such as graphs (e.g., Kleimuntz & Schkade, 1993). Importantly, the relative ease of using different kinds of information when evaluating graphs rather than number lists may influence the strategy participants employ to combine hedonic information in Varey and Kahneman's (1992) rating task. For example, numerical data points are more difficult for participants to extract when evaluating graphical rather than numerical displays (Jarvenpaa & Dickson, 1988; see also Simkin & Hastie, 1987). Increasing the difficulty of extracting individual data points via graphical formats may lower the likelihood that participants focus on individual moments of painful experiences, such as peak and end (relative to number lists). Instead, participants may be more likely to shift their attention to the total amount of pain represented by the graph (i.e., estimate the relative area of the graph occupied by pain ratings, which can be tantamount to an adding strategy).

Indeed, research suggests that participants may be less likely to focus on individual moments when evaluating graphs. When a great deal of information is presented (as it is in Varey & Kahneman, 1992), graphs rather than numerical representations of information help convey simple messages to participants (Dickson, DeSanctis, & McBride, 1986). Similarly, studies have shown that graphs may be best for communicating quick summaries of data (Jarvenpaa & Dickson, 1988; see also Hammond, 1971, who showed graphs facilitate rapid learning). Although these findings do not preclude participants from averaging ratings when presented with graphs, graphical representations may increase the opportunity for participants to focus attention on the experience in its entirety (i.e., total pain), rather than individual moments.

To make this *format-algorithm compatibility hypothesis* more intuitive, compare Fig. 1a and b. For the number lists in Fig. 1a, averaging the numbers seems easier or more natural than adding them, especially if the list were to increase in length (compare Fig. 1a, i. and ii.). In Fig. 1b, the same hedonic experiences are represented in graphical form and, in this case, adding the numbers seems easier or more natural than averaging them. In particular, it seems easier to estimate the proportion of the total graphical display occupied by the hedonic ratings, which is tantamount to an adding strategy (at least as presented in Fig. 1b). If participants were to use the strategy that appears most natural in each case, duration neglect would occur for the number list and disappear for the graphs.

To test whether graphical representations might lead people to use adding rather than averaging strategies, in Experiment 1 we presented some participants with the ratings in list form, just as Varey and Kahneman did, and we presented others with the ratings in graphical form (histograms). If graphical rather than numerical formats alter how people combine hedonic information—without increasing attention to duration—that would not only support the format-algorithm hypothesis, it would also raise questions about the cause and generality of duration neglect in two ways. First, reducing or eliminating duration neglect in Varey and Kahneman's task would suggest that duration neglect may be limited to remembered experiences, not experiences without a memory component (which contradicts the mainstream view; see Fredrickson & Kahneman, 1993; Kahneman et al., 1993; Schreiber & Kahneman, 2000; Redelmeier et al., 2003). Such a finding would introduce the possibility that duration neglect could be eliminated, or debiased (see Fischoff, 1982, for an explanation of debiasing). For example, when making a prospective choice between two unpleasant medical treatments, presenting patients with graphical hedonic representations of those experiences could result in choices that incorporate, rather than neglect, duration.

Second, the assertion that people tend to combine peak and end moments when evaluating hedonic experiences was first supported using data from Varey and Kahneman's (1992) hypothetical rating task (see also Fredrickson & Kahneman, 1993). If their findings were eliminated simply by changing the presentation format, that would bring into question the predominance of peak-ended behavior. Indeed, subsequent research has suggested that the combination of peak and end hedonic moments does not always best describe participants' evaluations of aversive experiences (e.g., Fredrickson & Kahneman, 1993). Furthermore, the ability to generalize peak-ended behavior to medical studies involving colonoscopies (Redelmeier & Kahneman, 1996; Redelmeier et al., 2003) has been disputed, primarily because the majority of patients were under the influence of anesthesia (Ariely et al., 2000).

Experiment 1 tests the format-algorithm hypothesis by examining whether graphs rather than number lists can make duration neglect disappear without increasing attention to episode duration. Two additional studies further explore the implications of the format-algorithm compatibility hypothesis on duration neglect. Experiment 2 investigates whether graphical formats facilitate adding (i.e., summing hedonic ratings) or area estimation (i.e., estimating the proportion of the total display occupied by hedonic ratings) so as to better understand what features of graphs cause participants to incorporate duration in their overall evaluations—which would be particularly important to debiasing efforts. Finally, Experiment 3 examines whether format influences the perceived ease of averaging, adding, or area estimation in a manner predicted by the format-algorithm compatibility hypothesis.

Experiment 1

Method

Participants

Participants were 113 University of California, San Diego (UCSD) students (35% males and mean age of 20 years), who received

Table 1
Stimuli used in Experiments 1, 2, and 3

Range	Duration		
	15 min	25 min	35 min
Ascending series:			
2–6	{2,4,6}	{2,3,4,5,6}	{2,2.67,3.33,4,4.67,5.33,6}
4–8	{4,6,8}	{4,5,6,7,8}	{4,4.67,5.33,6,6.67,7.33,8}
4–6	{4,5,6}	{4,4.5,5,5.5,6}	{4,4.33,4.67,5,5.33,5.67,6}
2–8	{2,5,8}	{2,3.5,5,6.5,8}	{2,3,4,5,6,7,8}
Descending series were the above sequences in reverse			
Additional series:			
30 min	{2,2,4,4,6,6} {6,6,4,4,2,2}	50 min	{7.5,7.5,8,8,8.5,9,9,9.5,9.5}
20 min	{2,5,8,4} {8,5,2,4}		{6,6,7,7,8,8,9,9,10,10}
			{5,5,6,6,7,7,8,8,9,9}
			{4,4,5,5,6,6,7,7,8,8}
Descending series for the 50 min sequences were the above sequences in reverse			

partial credit in psychology courses. They were randomly assigned to either the numerical (n = 37), full graphical (n = 38), or modified numerical (n = 38) format.

Procedure

All participants completed the survey in a laboratory setting. As in Varey and Kahneman's (1992) task, participants were asked to evaluate the uncomfortable experiences of hypothetical students who had participated in a series of experiments involving some time in an uncomfortable state (e.g., sat in a vibrating room, were exposed to loud drilling and hissing noises, stood in an uncomfortable position). Participants were told that each hypothetical student made a rating every 5 min of the discomfort they were feeling at that moment, using a scale from 0 to 10, where 0 represented no discomfort at all and 10 was an almost unbearable level of discomfort.

The participants were then instructed to evaluate the overall discomfort ratings made in each experiment on a scale from 0 to 100, where 0 was not bad at all and 100 was extremely bad. Before starting the evaluation process, participants were asked to look through the entire booklet of stimuli. (Full instructions are presented in the Appendix.) Instructions did not differ across conditions, with the exception of the sample stimulus presented at the end of the instructions, and a brief, one-sentence description of that sample stimulus. For example, the description for the numerical stimulus read "Note that each *line* represents a rating for one 5-minute interval", whereas the description for the graphical stimulus read "Note that each *column (vertical bar)* represents a rating for one 5-minute interval" (emphasis added). With the exception of the sample stimulus and the brief description of that stimulus, instructions were identical to Varey and Kahneman (1992); instructions were taken from their Appendix 3.

Design

Each participant evaluated 36 uncomfortable experiences (see Table 1). About half of the participants were presented with the 36 experiences in a predetermined random order, and half were presented with the reverse order. As in Varey and Kahneman's (1992) original task, 24 experiences (or sequences) comprised a within-participants factorial design with two trends (increasing and decreasing) × three episode durations (15, 25, and 35 min) × four intensity ranges (2–6, 2–8, 4–6, and 4–8). The 12 additional sequences were categorized as follows: four for planned comparisons ({2,5,8,4} and {8,5,2,4}; {2,2,4,4,6,6} and {6,6,4,4,2,2}) and eight for the purpose of increasing response-scale compatibility with adding strategies (see Table 1; all eight experiences were 50 min long).[1]

Uncomfortable experiences were presented in three different formats: two numerical and one graphical.[2] The first numerical format was intended to be identical to that of Varey and Kahneman's original study (Fig. 1a), and consisted of 2 columns with 3–10 rows (experiences ranged from 15 to 50 min). The 'full' graphical format consisted of histograms (Fig. 1b) that contained 10 columns (for a possible 50 min of discomfort), with 3–10 vertical bars shaded (a shaded vertical bar represented the magnitude of the rating for that particular 5-min period). Shaded vertical bars were divided into squares, each square representing one pain rating increment. If participants employed an area estimation strategy using the benchmark of 100 possible rating points for each experience (regardless of duration), results would be indistinguishable from that of adding.

Since the full graphical format could bring attention to maximum possible duration by always presenting 10 vertical spaces (corresponding to 50 min) for every experience, a 'modified' numerical format (Fig. 1c) was also included. The modified numerical format was intended to be more comparable to the full graphical format by including ten (horizontal) outlined spaces for each experience so that the effect of highlighting (maximum possible) duration could be evaluated. Like the graphs, sometimes these spaces were occupied by ratings and sometimes they were left empty. For instance, if the experience was 15 min long, 7 spaces would remain unoccupied. If the modified numerical format resulted in the same amount of duration neglect as the numerical format, an effect of duration in the full graphical format could not be explained by attention to maximum possible duration.

[1] The eight 50 min experiences were not used in Varey and Kahneman's (1992) original rating task, but were included in the present study to increase the maximum possible duration of Varey and Kahneman's stimuli from 35 to 50 min. By increasing maximum possible duration, the total possible pain was 100 (which is the maximum of the '0–100' response scale) instead of 70, presumably giving participants every opportunity to add ratings by making the stimuli more compatible with adding strategies (i.e., simple sum of ratings for numerical stimuli, and estimation of the proportion of the total display area occupied by ratings for graphical stimuli). One alternative to this method would have been to narrow the response scale from '0–100' to '0–70'.

[2] A second type of graph, pie charts, was also used. Analogous to the lists of numbers and histograms, there were 3–10 pie-pieces possible, and each piece indicated the magnitude of a rating at a particular 5-min interval. Results for the pie chart stimuli did not qualitatively differ from the histogram stimuli and, for the sake of exposition, we do not report them. Eliminating these results does not change our account in any way. Results for this condition are available from the first author.

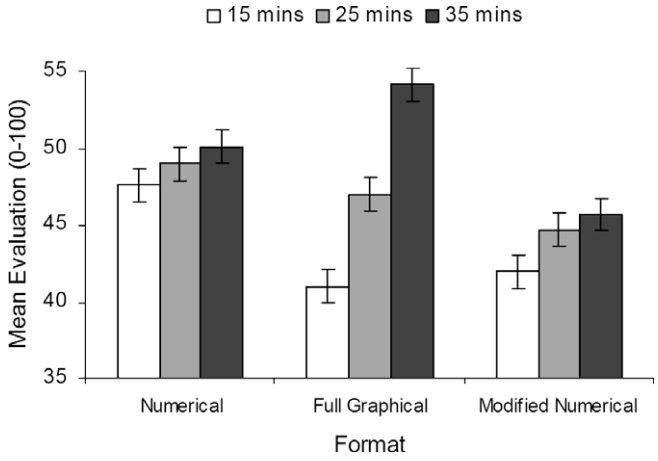


Fig. 2. Evaluation means at each duration level by format from Experiment 1. Standard error bars are shown.

Table 2
Percentage of participants using each utility combination strategy by format in Experiment 1 according to the (1) winner-takes-all analysis and (2) positive beta analysis

Format	Utility combination strategy			
	Adding (%)	Averaging (%)	Peak-ended (%)	Other (%)
Numerical				
Winner-takes-all	5	59	19	16
Positive beta	11	65	30	16
Full graphical				
Winner-takes-all	37	21	34	8
Positive beta	46	21	55	8
Modified numerical				
Winner-takes-all	8	29	58	5
Positive beta	16	39	61	5

Results

Duration neglect analysis

Comparing levels of duration neglect between each format is of primary interest in the current experiment. Thus, a 3 (Format: numerical, full graphical, modified numerical) × 3 (Duration: 15, 25, 35 min) mixed-model ANOVA on global evaluations for the 24 stimuli shown at the top of Table 1 was performed, where Duration was within-participants and Format was between-participants. The analysis indicated that two effects were significant: the main effect of Duration, $F(2,220) = 27.2$, $p < .001$, and the Format × Duration interaction, $F(4,220) = 7.5$, $p < .001$. In terms of the former effect, evaluations at 35 min were highest, followed by 25 and 15 min (Ms = 50, 47, and 44, respectively). Of primary importance, the Format × Duration interaction showed that participants largely neglected duration for both the numerical and modified numerical formats, but they did not neglect it for the full graphical format (Fig. 2). Pre-planned contrasts between mean responses at each duration level yielded no significant differences ($p > .05$) for the numerical format, a difference between only the 15 and 35 min durations for the modified numerical format ($p = .02$), while for the full graphical format, there were significant differences between each duration level ($p < .001$).

Recall that the modified numerical format was created to emulate the full graphical format: both formats highlighted maximum possible duration. In this way, it could be determined whether highlighting duration affected global evaluations. If the modified numerical format showed an equivalent level of duration neglect relative to the numerical format, then the effect of duration in the full graphical format could not be explained by attention to (maximum possible) duration.

To test this, a separate 2 (Format: numerical, modified numerical) × 3 (Duration: 15, 25, 35 min) mixed model ANOVA on global evaluations for the 24 stimuli shown at the top of Table 1 was performed. Only a small, but reliable, main effect of Duration was significant, $F(2,146) = 11.3$, $p < .001$: evaluations were lowest at 15 min, followed by 25 and 35 min (Ms = 45, 47, and 48, respectively). Importantly, there was no Format × Duration interaction ($p = .59$) indicating that levels of duration neglect for both numerical formats were very similar, even though the modified numerical format highlighted (maximum possible) duration for each experience.[3]

Format-algorithm compatibility analysis

In an attempt to uncover what strategies participants used to arrive at their global evaluations, we regressed 28 of participants' 36 responses on adding, averaging, and peak-ended strategies (the 8 filler stimuli were not included in this analysis). The adding strategy was defined as the sum of an episode's hedonic ratings; the averaging strategy was defined as the sum of an episode's hedonic ratings divided by the total number of hedonic ratings; and the peak-ended strategy was defined as the sum of an episode's most aversive and final hedonic ratings divided by two. For example, for the 15 min episode of increasing discomfort {2,5,8} the adding strategy would equate to an evaluation of 15 (2 + 5 + 8), the averaging strategy 5 ([2 + 5 + 8]/3), and peak-ended strategy 8 ([8 + 8]/2).

Letting R represent a participant's response, and X, Y, and Z represent the sum, average, and peak-ended average, respectively, we fit a linear model of the form $R = a + b\mathrm{X} + c\mathrm{Y} + d\mathrm{Z}$. This linear model enabled us to identify whether the participant added, averaged, and/or used a peak-ended strategy by testing the hypothesis that b, c, or d was zero ($p < .05$) via a partial R-squares analysis (if b, c, or d, was significantly greater than zero, that would mean that the participant added, averaged, or used a peak-ended strategy, respectively). A partial R-squares analysis measures the marginal contribution of each strategy when all other strategies are already included in the model (i.e., it excludes the variance that is not unique to a particular strategy). If parameters were either 0 or negative for *all* strategies, participants were categorized as "other".

Table 2 shows the proportion of participants whose dominant utility combination strategy was adding, averaging, or peak-ended (the *winner-takes-all* analysis). A participant's dominant strategy was defined as the strategy that explained the largest portion of that participant's response variance relative to the other strategies. For example, if the adding strategy explained the most response variance for a particular participant, that participant would be included in the "Adding" category of Table 2 for the winter-takes-all analysis. Additionally, the table indicates the proportion of participants that used any of the three strategies (*positive beta* analysis). For example, if a participant had a significant positive b, c, and d (i.e., beta) value for all three strategies, that participant would be included in the "Adding", "Averaging", and "Peak-ended" catego-

[3] It is also possible that the effect of duration was *not* significantly greater for the full graphical relative to the modified numerical format, which would be problematic for the format-algorithm compatibility hypothesis. To evaluate this possibility, a separate 2 (Format: graphical, modified numerical) × 3 (Duration: 15, 25, 35 min) mixed model ANOVA on global evaluations for the 24 stimuli shown at the top of Table 1 was also performed. Both the main effect of Duration, $F(2,148) = 22.7$, $p < .001$, and the Format × Duration interaction, $F(2,148) = 7.5$, $p < .001$, were significant, showing that global evaluations increased with duration length (Ms = 42, 46, and 50, at 15, 25, and 35 min, respectively), and that the increase was more pronounced for the full graphical relative to the modified numerical format (i.e., the effect of duration was greater for the full graphical format; see Fig. 2).

ries of Table 2 for the positive beta analysis—without regard for the relative amount of response variance each strategy explained.

The winner-takes-all analysis showed that for the numerical format, the majority of participants were best explained by an *averaging* strategy, whereas for the full graphical format, the modal participant was best explained by an *adding* strategy. The positive beta analysis yielded somewhat similar qualitative results, although it appears that the winner-takes-all analysis deemphasized the degree to which a peak-ended strategy played a role in participants' evaluations for graphs.

Interestingly, the numerical and modified numerical formats showed somewhat different patterns of behavior: while the modal participant used a simple averaging strategy to combine ratings for the numerical format, the modal participant used a peak-ended averaging strategy for the modified numerical format. It is likely that differences were due to random variation in averaging strategies. However, there is also the possibility that alterations to the modified numerical format caused an increased focus on peak and end ratings—which would only lend credence to the hypothesis that Varey and Kahneman's (1992) results are format dependent. In any case, it appears that although peak-ended behavior is not a general rule in combining hedonic ratings in this task, it can still play an important role in participants' rating combination strategies.

Group-level comparisons with Varey and Kahneman (1992)

Based on group-level analyses, Fredrickson and Kahneman (1993) concluded that a weighted average of peak and end ratings best characterized people's evaluation strategies in Varey and Kahneman's (1992) Experiment 2. The *individual-level* analysis of the current numerical format—which, to our knowledge, was identical to Varey and Kahneman's format—suggested that a simple averaging strategy best described people's evaluations. Thus, it was possible that the numerical format failed to replicate Varey and Kahneman's group-level results. To explore this issue, the 28 evaluation means in Varey and Kahneman's task (taken from their Exhibit 3) were correlated with the equivalent 28 evaluation means elicited from the current numerical format (again, the 8 filler stimuli were not used). The high correlation between the two sets of means (.89) indicated that we successfully replicated Varey and Kahneman's results. The successful replication implies that a group-level analysis of means may be misleading when inferring participants' strategies (at least in this case). To investigate this further, we correlated peak-ended and averaging strategies with the group-level means of the current numerical format. The correlations for averaging and peak-ended strategies were nearly identical (.87 vs. .84, respectively), even though at the individual-level, over 3 times the participants were shown to predominately average rather than combine peak and end ratings. Group-level analyses make the peak-ended strategy appear more likely than the individual-level analyses suggest.

Discussion

The primary goal of Experiment 1 was to examine whether changing hedonic ratings from a numerical to a graphical format would induce an adding strategy (i.e., estimation of the proportion of the total display occupied by the ratings), effectively eliminating duration neglect. Indeed, we found that duration was not neglected when participants evaluated graphs, but it was neglected when participants evaluated lists of numbers. The most common rating combination strategy in the 'full' graphical format was adding, while a majority of participants appeared to average ratings in the numerical format.

Although the full graphical format appeared to highlight the maximum possible duration of each experience, whereas the numerical format did not, the results cannot be explained by this difference: a modified numerical format—which was designed to emulate the full graphical format by highlighting maximum possible duration—showed no differences in levels of duration neglect relative to the numerical format. If attention to duration (caused by highlighting maximum possible duration) was responsible for the effect of duration for graphs, we should have also observed a larger effect of duration in the modified numerical relative to the numerical format. Format-algorithm compatibility, not attention to duration, appears to explain the pattern of results.

Furthermore, recall that Varey and Kahneman (1992) concluded from a *group-level* analysis that a weighted average emphasizing the peak and end moments of an episode best characterized participants' evaluation strategies. We found that both averaging and peak-ended strategies were similarly (and highly) correlated with group-level responses for the current numerical format. However, the *individual-level* analysis of the numerical format revealed that a simple average of all ratings best described 59% of participants' behavior and that only 19% of these participants were best described by a peak-ended average. Why did group- and individual-level analyses differ? It appears that a minority of participants had a disproportionately large effect at the group-level: when we removed the 19% of participants who were predominately "peak-ended" from the numerical format, group-level means were almost perfectly correlated with simple averaging (.95) and much less correlated with a peak-ended average (.63; the correlation between simple averaging and a peak-ended average is .55).

Our results are consistent with other research showing that group-level analyses can be misleading (e.g., Pashler, 1998; Rickard, 2004) and lend credence to the hypothesis that format-algorithm compatibility, not attention to peak and end ratings, was (at least in part) the cause of duration neglect in Varey and Kahneman's (1992) rating task: the numerical format presented to participants was more compatible with averaging strategies that often included all ratings, not just peak and end.

Experiment 2

The format-algorithm compatibility hypothesis posits that particular rating combination strategies are more or less compatible with particular formats. In Experiment 1, we hypothesized and found that the full graphical format is more compatible with adding strategies. Experiment 2 examines which features of graphs cause participants to utilize adding strategies, which not only has theoretical, but practical implications: if graphical formats are used to debias duration neglect, knowing what features of the format are important to eliminating the effect is critical.

Recall that the full graphical format in Experiment 1 (Fig. 1b) was hypothesized to promote an area estimation strategy—i.e., estimation of the proportion of the total display occupied by the ratings—rather than summing the ratings per se. Such an area estimation strategy would be consistent with adding because the area occupied by the ratings is always evaluated against the same benchmark (i.e., the same total display area of 10 vertical bars), regardless of episode duration (compare Fig. 1b, i. and ii.). To the extent that area estimation rather than adding per se occurred for the full graphical format, cutting-off the *x*-axis at the end of the episode should make duration neglect reappear. To test this, a modified graphical format was created. Rather than always containing the same total display area of 10 vertical bars as with the full graphical format in Experiment 1, for the modified graphical format, the total display area corresponded to the length of the experience (compare Fig. 1d, i. and ii.). In other words, graphs representing durations of 15 min contained 3 vertical bars; graphs representing durations of 25 min contained 5 vertical bars; graphs

representing durations of 35 min contained 7 vertical bars; and so on. If participants estimated the proportion of the total display occupied by the ratings for Experiment 1's full graphical format, cutting-off the *x*-axis at the end of the experience should make a difference: evaluating the area occupied by the ratings against the changing (rather than the constant) benchmark should cause results to be analogous to averaging rather than adding (i.e., duration should be neglected). However, if participants added ratings in Experiment 1's full graphical format, cutting-off graphs upon termination of the experience should matter little (i.e., duration should not be neglected). The upshot is that if graphs are more compatible with adding per se, changing the features of the graph should not impact the effect of duration found in Experiment 1, but if graphs are more compatible with area estimation, changing the features will impact the effect of duration.

However, notice that the modified graphical format increases in length as duration increases (compare Fig. 1d, i. and ii.). This may increase attention to relative episode duration, thereby causing participants to incorporate duration in their global evaluations, regardless of the rating combination strategy participants would typically use for graphs. Thus, a second modified graphical format was created that held constant the length of the *x*-axis for all graphs (compare Fig. 1e, i. and ii.). In this way, the effect of attention to duration in the modified graphical format (Fig. 1d)—which highlighted relative episode duration—could be evaluated by comparing it to the modified-fixed-axis graphical format (Fig. 1e)—which did not highlight relative episode duration. If the modified and modified-fixed-axis graphical formats yield similar effects of duration, then increased attention to duration would seem to play little role in participants' rating combination strategies, thereby lending further credence to the format-algorithm compatibility hypothesis.

Method

Participants

Participants were 120 UCSD students (25% males and mean age of 21 years), who received partial credit in psychology courses. They were randomly assigned to either the full graphical ($n = 40$), modified graphical ($n = 40$), or modified-fixed-axis graphical ($n = 40$) formats. One participant in the modified-fixed-axis graphical format gave an uninterpretable response and another in the modified graphical format skipped one rating. Both participants were excluded from the analyses, leaving 39 in these two formats.

Procedure

The procedure was identical to that of Experiment 1.

Design

The design was identical to that of Experiment 1, with the exception that uncomfortable experiences were presented in three different formats, all graphical. The full graphical format described in Experiment 1 remained the same (Fig. 1b) and two modified graphical formats were added: the modified graphical (Fig. 1d) and modified fixed-axis graphical (Fig. 1e) formats.

Results

Duration neglect analysis

As in Experiment 1, comparing levels of duration neglect between each format is of primary interest in the current experiment. Thus, a 3 (Format: full graphical, modified graphical, modified-fixed-axis graphical) × 3 (Duration: 15, 25, 35 min) mixed model ANOVA on global evaluations for the 24 stimuli in Table 1 was performed, where Duration was within-participants and Format was between-participants. There were main effects of Format, $F(2,115) = 4.9$, $p = .009$, and Duration, $F(2,230) = 60.7$, $p < .001$, as well as a Format × Duration interaction, $F(4,230) = 4.6$, $p = .001$.

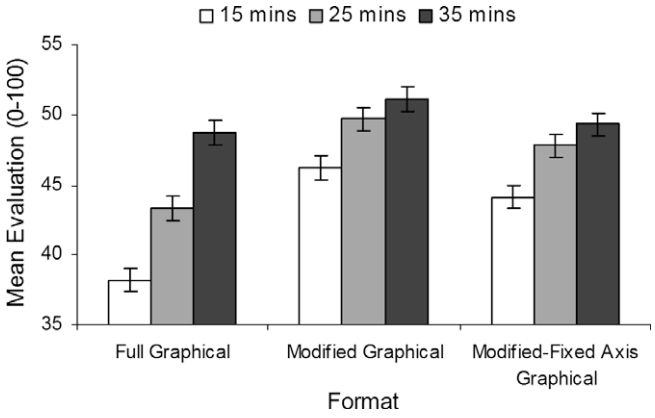


Fig. 3. Evaluation means at each duration level by format from Experiment 2. Standard error bars are shown.

The main effects were of little importance and are described briefly. For the main effect of Format, evaluations for the full graphical format were the lowest, followed by the modified and modified-fixed-axis formats (*M*s = 43, 47, and 49, respectively). The main effect of Duration showed that evaluations at 35 min were highest, followed by 25 and 15 min (*M*s = 50, 47, 43, respectively). Most importantly, the Format × Duration interaction indicated that there were different effects of duration on global evaluations by format (Fig. 3): whereas participants partly neglected duration for both the modified and modified-fixed-axis graphical formats, they did not neglect it for the full graphical format. Pre-planned contrasts between participants' mean responses at each duration level yielded significant differences ($p < .001$) for both the modified and modified-fixed-axis graphical formats at episode durations of 25 and 35 min, as well as 15 and 35 min, while for the full graphical format, there were significant differences between each duration level ($p < .001$).

Recall that the modified graphical format highlighted relative episode duration to a larger degree than the other two graphical formats. If format-algorithm compatibility, rather than lack of attention to duration, drove the results, then levels of duration neglect in the modified graphical format should be similar to that of the modified-fixed-axis graphical format. To evaluate this conjecture, we performed a 2 (Format: modified graphical, modified-fixed-axis graphical) × 3 (Duration: 15, 25, 35 min) mixed model ANOVA on global evaluations for the 24 stimuli in Table 1. Only the main effect of Duration was significant, $F(2,152) = 27.6$, $p < .001$, indicating a small, but reliable effect of duration (*M*s = 45, 49, and 50 for 15 min, 25 min, and 35 min episodes, respectively). Of primary importance, the absence of the Format × Duration interaction ($p = .99$) confirmed that duration was partly neglected (at a similar level) for both the modified graphical formats, further supporting the format-algorithm hypothesis.[4]

[4] It is also possible that the effect of duration was *not* significantly greater for the full graphical relative to the modified graphical format. To evaluate this possibility, a separate 2 (Format: full graphical, modified graphical) × 3 (Duration: 15, 25, 35 min) mixed model ANOVA on global evaluations for the 24 stimuli shown at the top of Table 1 was also performed. Both the main effect of Duration, $F(2,154) = 49.4$, $p < .001$, and the Format × Duration interaction, $F(2,154) = 6.2$, $p = .003$, were significant, showing that global evaluations increased with duration length (*M*s = 41, 46, 49, at 15, 25, and 35 min, respectively), and that the increase was more pronounced for the full graphical relative to the modified graphical format (i.e., the effect of duration was greater for the full graphical format; see Fig. 3).

Table 3
Percentage of participants using each utility combination strategy by format in Experiment 2 according to the (1) winner-takes-all analysis and (2) positive beta analysis

Format	Utility combination strategy			
	Adding (%)	Averaging (%)	Peak-ended (%)	Other (%)
Full graphical				
Winner-takes-all	33	28	38	3
Positive beta	50	30	50	3
Modified graphical				
Winner-takes-all	13	44	33	10
Positive beta	18	51	44	10
Modified-fixed-axis graphical				
Winner-takes-all	18	38	38	5
Positive beta	31	41	49	5

Format-algorithm compatibility analysis

As in Experiment 1, we regressed 28 of participants' 36 responses on adding, averaging, and peak-ended strategies (Table 3). The winner-takes-all analysis showed that adding strategies were most common for the full graphical format, followed by the modified-fixed-axis and modified graphical formats. Averaging showed the reverse pattern: the largest percentage of participants employed an averaging strategy for the modified graphical format, followed by the modified-fixed-axis and then full graphical formats. The positive beta analysis showed that half of participants used an adding strategy when combining ratings for the full graphical format—about twice the number of participants using an adding strategy for the modified and modified-fixed-axis graphical formats. Finally, although peak-ended strategies played a large role in participants' responses, for each format, the majority of participants appeared to use an adding or averaging strategy that incorporated *all*, rather than just peak and end, ratings to form global evaluations of experiences.

Discussion

Further evaluation of format-algorithm compatibility indicated that participants did not estimate sums of ratings for graphs, but instead utilized an area estimation strategy, whereby the proportion of the total graphical display occupied by the ratings was estimated. In particular, there was a relatively large effect of duration for the full graphical format, where area estimation would result in adding. Conversely, duration had a small effect on both the modified (Fig. 1d) and modified-fixed-axis (Fig. 1e) graphical formats, where area estimation would result in averaging. It is noteworthy that duration was neglected for the modified graphical format although the format highlighted relative episode duration to a greater degree than the other two graphical formats. Indeed, individual-level analyses showed that relative to the full graphical format, about 50% fewer participants used an adding strategy while about 50% more participants appeared to average for the modified and modified-fixed axis graphical formats. These results illustrate that it is not just the format itself, but the features of the format, that matters when evaluating the efficacy of graphs in reducing levels of duration neglect.

Experiment 3

Taken together, Experiments 1 and 2 suggest that area estimation, rather than averaging or adding, is easier or more natural for graphs. However, for lists of numbers, the results suggest that averaging is easier or more natural than adding. The third and final experiment tested in a more direct fashion the validity of the format-algorithm compatibility hypothesis by asking participants to report their perceptions of the ease and naturalness of estimating either (1) averages vs. sums for lists of numbers, (2) averages vs. sums for graphs, or (3) averages vs. area for graphs. It was predicted that participants would report that averaging was easier and more natural when averaging and summing lists of numbers and graphs. However, when averaging and estimating area for graphs, it was predicted that participants would find area estimation easier and more natural.

Method

Participants

Participants were 94 UCSD students (31% males and mean age of 20 years), who received partial credit in psychology courses. They were randomly assigned to either compare averaging and adding strategies for lists of numbers ($n = 32$), averaging and adding strategies for graphs ($n = 32$), or averaging and area estimation strategies for graphs ($n = 30$). One participant in the number list condition failed to complete the experiment and therefore was not included, leaving 31 in that condition.

Procedure

All participants completed a computer task and a survey in a laboratory setting. For the computer task, one group of participants was asked to perform calculations on number lists, while two other groups were asked to perform calculations on graphs (stimuli were identical to Fig. 1b and c except that duration information was removed). Each participant performed calculations on the 36 stimuli in Table 1. All stimuli were presented on computer monitors and calculations were recorded by participants on a response sheet.

For lists of numbers and one of the graphical conditions, participants were asked to estimate averages and sums. For the other graphical condition, participants were asked to estimate averages and area. Sample calculations were provided using either numbers or graphs (depending on the condition), and were as follows: for averages, participants were shown an example where values were summed, and then divided by the number of values in the sequence; for sums, participants were simply shown an example where values were summed; for area, the sample calculation showed a sum of the values, divided by the maximum total possible value (which was always 100 for this set of stimuli). Participants were informed that a stimulus would be presented for 1 s, after which time they would have 2 s to record their calculation before being given an additional 2 s to prepare for the next stimulus.

After participants viewed and performed calculations on all 36 stimuli, they were asked to complete a brief paper-based questionnaire containing 5 questions. Of the 5 questions, one was a filler question (e.g., "In your everyday life do you think you add or average lists of numbers more often?"), 2 questions were the primary dependent measures of interest, and 2 questions were of secondary interest. The 2 questions of primary interest were related to ease and naturalness:

(1) In this experiment, did you find it *easier* to add [calculate area] or average?
(2) In this experiment, did it feel *more natural* to add [calculate area] or average?

The 2 questions of secondary interest were related to accuracy and precision, and were included to examine whether these factors might also be related to format-algorithm compatibility:

(3) In this experiment do you think you got the *right answer* more often when adding [calculating area] or averaging?
(4) In this experiment do you think you made *big errors* more often when adding [calculating area] or averaging?

Participants answered questions on a scale of −3 to 3, where '−3' and '3' indicated strong agreement for a particular strategy (i.e., adding, averaging, or area estimation) and '0' indicated indifference between strategies.

Design

About half of the participants were presented with the 36 experiences in a predetermined random order, and half were presented with the reverse order. Furthermore, half of participants were randomly assigned to average 18 of the experiences and sum [estimate area of] the remaining 18 experiences, while for the other half of participants, strategy assignment was reversed. The averaging and adding [area estimation] strategies were intermixed, not blocked.

Results

Ease and naturalness

Of primary interest is participants' perception of the relative ease and naturalness of particular strategies for numerical and graphical formats. Because participants' responses to the "ease" and "naturalness" questions were highly correlated ($r = .89$), we averaged them for each participant (for the sake of exposition).

Recall that participants answered questions on a scale of −3 to 3, where '−3' and '3' indicated strong agreement for a particular strategy and '0' indicated indifference between strategies. For each of the three conditions, a preference for averaging was coded as a positive number ('1' to '3')—since it was the common competing strategy across all three conditions—and a preference for adding [area estimation] was coded as a negative number ('−1' to '−3').

To examine the relative ease and naturalness of averaging compared to adding for lists of numbers and graphs, and averaging compared to estimating area for graphs, a between-participant one-way ANOVA (Condition: numerical [averaging vs. add]; Graphical [averaging vs. add]; Graphical [averaging vs. area estimation]) was performed on participants' combined "ease" and "naturalness" responses. The main effect of Condition was significant, $F(2,90) = 6.2$, $p = .003$. Participants perceived averaging as easier and more natural when asked to average and add for lists of numbers ($M = .8$) and graphs ($M = 1.7$), but averaging was relatively more difficult and less natural when asked to average and estimate area for graphs ($M = -.1$).

Accuracy and precision

Also of interest is participants' perception of the relative accuracy and precision for particular strategies. As with "ease" and "naturalness", participants' responses to the "accuracy" and "precision" questions were highly correlated ($r = .84$), and were therefore averaged for the sake of exposition (after first multiplying the precision responses by −1, since original responses were related to levels of perceived *im*precision, rather than precision).

To examine the (perceived) relative accuracy and precision of averaging compared to adding for lists of numbers and graphs, and averaging compared to estimating area for graphs, a between-participant one-way ANOVA (Condition: numerical [averaging vs. add]; Graphical [averaging vs. add]; Graphical [averaging vs. area estimation]) was performed, but this time on participants' combined responses to the "accuracy" and "precision" questions. Results indicated that the main effect of condition was significant, $F(2,90) = 3.2$, $p = .04$. Similar to results for ease and naturalness, findings show that participants perceived averaging as more accurate and precise when asked to average and add for lists of numbers ($M = .9$) and graphs ($M = 1.2$), but that averaging was relatively inaccurate and imprecise when asked to average and estimate area for graphs ($M = 0$). (Responses for ease/naturalness and accuracy/precision were highly correlated, $r = .85$.)

Perceived vs. actual accuracy and precision

We also wanted to know whether *actual* accuracy and precision were related to *perceptions* of accuracy and precision. Correlations were calculated between participants' (1) perceived relative correct answers and actual relative correct answers[5] and (2) perceived relative errors and actual relative errors.[6] For accuracy and errors, the correlations were .43 ($p < .001$) and.27 ($p < .008$), respectively, indicating that participants had reasonably good insight into their accuracy and precision levels.

Discussion

Participants reported that averaging was easier and more natural than adding for lists of numbers and graphs, but that area estimation was at least as easy and natural relative to averaging for graphs. These results provide evidence for the format-algorithm hypothesis and are consistent with other research showing that different formats may facilitate different strategies because those strategies are perceived as relatively easier or more natural to use (e.g., Kleimuntz & Schkade, 1993). These findings also support the idea that format-algorithm compatibility, rather than lack of attention to duration, contributed to the duration neglect in Varey and Kahneman (1992): it is relatively easier and more natural to average rather than add lists of numbers.

Similarly, the results also showed that participants perceived averaging to be more accurate and precise than adding for lists of numbers and graphs, but that they perceived area estimation to be at least as accurate and precise as averages for graphs. This suggests that perceptions of accuracy and precision may also contribute to format-algorithm compatibility (see Payne et al., 1993, for a related discussion). It is also interesting that participants had reasonably good insight into their relative accuracy and precision: there were moderately positive correlations between perceived and actual accuracy and precision for the different strategies.

General discussion

Varey and Kahneman's (1992) original demonstration of duration neglect is, at least in part, due to format presentation rather than attention to peak and end ratings. Experiment 1 showed that changing representations of hedonic experiences from a numerical to a graphical format eliminated duration neglect without increasing attention to duration. Format-algorithm compatibility appears to explain this pattern of results: in some cases it is easier or more natural to estimate averages for lists of numbers and to estimate area for graphical representations of numbers. Experiment 2 supported this notion by showing

[5] To calculate actual relative accuracy for each participant, the number of correct responses for each of the two competing strategies was determined, and then divided by the total number of possible correct responses for each strategy (i.e., number correct out of 18). The proportion of correct responses for adding/area estimation strategies was then subtracted from the proportion of correct responses for averaging strategies.

[6] To calculate actual relative errors, mean proportion error for the two competing strategies was first determined for each participant. Specifically, for each of the participant's 36 responses, the absolute distance of each response from the correct answer was calculated (i.e., |participant response − correct response|), and then divided by 10 for averaging and 100 for adding as well as area estimation (participant responses were converted from a 1 point scale to a 100 point scale for area estimation to aid analyses). The proportion error of the 18 responses for each of the two competing strategies was then averaged for each participant. Each participant's mean proportion error for adding/area estimation was then subtracted from the mean proportion error for averaging.

that, even for graphical formats, participants neglected duration when the format made area estimation consistent with averaging rather than adding. Finally, Experiment 3 directly tested the format-algorithm hypothesis by asking participants which strategies were easier or more naturally employed for number lists and graphs. Participants perceived averaging to be easier and more natural than adding for lists of numbers, but they perceived area estimation to be at least as easy and natural relative to averaging for graphs. Participants were also asked about perceptions of accuracy and precision, which showed a similar pattern of results to those of ease and naturalness. Along with ease and naturalness (e.g., Kleimuntz & Schkade, 1993), perceived accuracy and precision may play an important role in determining the choice of strategy (e.g., Chu & Spires, 2003; Payne et al., 1993).

These findings fit into a larger literature showing that format presentation can influence how information is integrated. For example, some researchers have shown that data presented as frequencies rather than probabilities can lead both undergraduates and experts to combine that data in substantively different ways (e.g., Gigerenzer & Hoffrage, 1995; see also Martignon, 2001). Similarly, others have shown that presenting potential risks graphically rather than numerically significantly alters risk taking behavior (Stone, Yates, & Parker, 1997; see also Chua, Yates, & Shah, 2006). In combination with the results of the current paper, such format dependence suggests caution when making generalizations about phenomena when format has the potential to significantly influence decision-making.

One might speculate that graphical formats induced a more deliberate, algorithm-based approach relative to the numerical formats. Borrowing from Sloman (2002), this would suggest that more controlled/rule-based (i.e., System 2) processes were responsible for the normative result in the graphical conditions (i.e., temporal integration), but that more automatic/associative (i.e., System 1) processes were responsible for the non-normative preferences in the numerical conditions (i.e., weighted average of prototypical moments; see also Kahneman & Frederick, 2002; Stanovich & West, 2000). Indeed, Kahneman and Frederick (2002) have (at least in part) attributed duration neglect to a "lazy" System 2 (see also Kahneman, 2003). However, we found that most participants reported estimating, not formally calculating their impressions of experiences, regardless of presentation format. In questionnaires administered after Experiment 1, participants were asked, "did you estimate (e.g., "eyeball") or formally calculate your answers?" 84% and 89% of participants evaluating the numerical and graphical formats, respectively, said that they "eyeballed" evaluations. These reports provide evidence that System 2 processes were not responsible for eliminating duration neglect in the graphical format. That is, it does not appear that participants generated an explicit algorithm to integrate ratings for the graphical format, but instead chose an estimation strategy most compatible with the stimuli (in this case, an area estimation strategy). The fact that the graphical stimuli efficiently (i.e., via System 1 processes) and effectively induced a normative strategy is consistent with earlier ideas related to "visual thinking", whereby images, rather than alternative constructs such as language, are said to best facilitate proper reasoning (see Arnheim, 1969).

Potential debiasing opportunities also arise from the graphical format's elimination of duration neglect (see Fischoff, 1982). For example, imagine that a patient underwent two different chemotherapy regimens, and the only difference between them was that the second regimen included a period of diminishing pain. Although previous duration neglect research would imply that the patient would choose to repeat the second regimen (e.g., Kahneman et al., 1993), if the patient were presented with the two hedonic experiences in graphical form, the patient's preference might reverse.

Our results suggest that evaluations of hedonic episodes that do and do not involve memory may not be as psychologically related as has been claimed (e.g., Frederickson, 2000; Kahneman, 2000b; Kahneman et al., 1997). Even when duration was neglected in the current study, people did not appear to only use prototype, or peak and end, moments when evaluating experiences. Instead our results showed that participants were often willing to use *all* hedonic moments. This result is particularly surprising because researchers have suggested that duration neglect is not only caused by memory constraints, but the privileged 'meaning' (e.g., Frederickson, 2000) of prototype moments (see also Ariely & Carmon, 2000, for a related explanation). If the special meaning of peak and end moments were the primary determinant of duration neglect in Varey and Kahneman's (1992) task, the majority of participants in our experiments should have weighed those moments more heavily than all other moments—even in the absence of a memory component (which they did not). Thus, peak and end evaluations may be a phenomenon primarily associated with remembered utility (e.g., "how much did I (dis)like that experience?"), where prototypical moments provide a 'meaningful' approximation, or snapshot, of the hedonic experience, thereby bypassing the need to store or recreate the entire experience in memory. But when a memory component is not involved, as in the present study, the need to rely on such heuristics appears to be largely mitigated.

However, researchers may even need to be cautious when claiming that people average peak and end moments when evaluating events from memory. Many such claims have been primarily based on group-level analyses (e.g., Redelmeier & Kahneman, 1996; Schreiber & Kahneman, 2000; Varey & Kahneman, 1992), which we showed can make a peak-ended strategy appear more likely than individual analyses suggest. (In fact, using an individual-level analysis, Fredrickson & Kahneman, 1993, found that a peak-ended strategy did not best explain participants' retrospective evaluations of aversive films.) In Experiment 1, we found that, at the group-level, evaluation patterns for the numerical format were consistent with both peak-ended and simple averages, but that individual-level analyses told a different story: more than 3 times the participants were best explained by a simple average relative to a peak-ended average. This suggests that if, for example, people actually use averaging models to update their impressions as experiences unfold over time, group-level analyses might nonetheless lead to the conclusion that a peak-ended strategy is common. This matters both theoretically and practically. Averaging and weighted (peak-ended) averaging models make different predictions regarding people's evaluations of hedonic experiences. For instance, imagine that, in accordance with previous prescriptions (e.g., Redelmeier & Kahneman, 1996; Redelmeier et al., 2003) a diminishing period of discomfort was added to an uncomfortable experience rated {2,5,8}. Now imagine that this diminishing period of discomfort was rated a '6'. When comparing experiences of {2,5,8} and {2,5,8,6} a "peak-ender" would prefer {2,5,8,6} (8 vs. 7), whereas an "averager" would prefer {2,5,8} (5 vs. 5.25). In other words, although adding a diminishing period of pain would make a "peak-ender" better off, it would make an "averager" *worse off*.

If peak-ended prescriptions have the potential to make people worse off, it is important to look more closely at individual-level, rather than (potentially misleading) group-level, results. For example, arthritis sufferers have been shown to exhibit peak-ended evaluative behavior using group-level analyses (Stone, Broderick, Kaell, DelesPaul, & Porter, 2000)—analyses which, after being revisited at the individual-level, may reveal that people are in fact

averaging all hedonic moments. It is particularly important to reexamine such assertions now that the idea of peak-ended-ness has become so prolific that even non-significant, suggestive, group-level analyses have been used to support a peak-ended evaluative model (in cancer research, see Jansen, Kievit, Nooij, & Stiggelbout, 2001).

Appendix

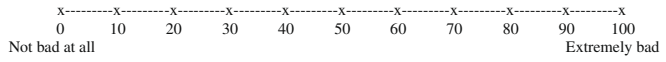
In this survey we are interested in people's intuitions about uncomfortable experiences. Students were paid to participate in a series of experiments. Each experiment involved some time in an uncomfortable state, such as sitting in a vibrating room, exposure to loud drilling or hissing noises, standing in an uncomfortable position, etc. The participants were told at the beginning of each experience how long it would last. Although the experimental conditions did not change in the course of a session, the subjective experience of discomfort often changed over time. The participants in each experiment made a rating every 5 min of the discomfort they were feeling *at that moment*. The last rating was made just before the end of the experiment. These ratings are on a scale from 0 to 10, as follows:

0 = no discomfort at all
10 = almost unbearable

In each of the following questions, you will be given the average discomfort ratings made in an experiment, and your task will be to provide an overall evaluation of the experience of a participant in that experiment.

In interpreting the discomfort ratings you should keep in mind that the participants served in a series of such experiments, and were highly trained in the use of discomfort scale. In particular, they were instructed to use the scale consistently, so that a rating of 5, for example, indicates the same level of subjective discomfort at the beginning, in the middle or at the end of an experimental session. You should also assume that the experiences have no after-effects of pain or discomfort—the participants return very quickly to a normal level of comfort.

For each set of these pain ratings, we would like you to provide a global evaluation of *how bad the overall experience is*, using a scale from 0 to 100, as follows:



Each experience is presented in the following way (*there is no need to evaluate this example*)[7]:

The ratings are:

after 5 min	9	
after 10 min	9	
after 15 min	10	
after 20 min	10	
after 25 min	8	Your global evaluation of the experience?
after 30 min	8	
after 35 min	9	
after 40 min	9	_____ (0–100)
after 45 min	1	
at end 50 min	5	

Note that each line represents a rating for one 5-minute interval. Thus, in the example above, the discomfort rating is 9 for the first 5-minute interval. The experiences you will evaluate have a maximum possible duration of ten 5-minute intervals (i.e., 50 min).

Please take a moment to look through the booklet before you begin. Then answer the questions in order, without looking back or ahead. [8]

References

Ariely, D., & Carmon, Z. (2000). Gestalt characteristics of experiences: The defining features of summarized events. *Journal of Behavioral Decision Making, 13*, 191–201.
Ariely, D., & Loewenstein, G. (2000). When does duration matter in judgment and decision making? *Journal of Experimental Psychology: General, 4*, 508–523.
Ariely, D., Kahneman, D., & Loewenstein, G. (2000). Joint comment on "When does duration matter in judgment and decision making?" (Ariely & Loewenstein, 2000). *Journal of Experimental Psychology: General, 129*, 524–529.
Arnheim, R. (1969). *Visual thinking*. Berkeley, CA: University of California Press.
Chu, P. C., & Spires, E. E. (2003). Perceptions of accuracy and effort of decision strategies. *Organizational Behavior and Human Decision Process, 91*, 203–214.
Chua, H. F., Yates, F. J., & Shah, P. (2006). Risk avoidance: Graphs versus numbers. *Memory & Cognition, 34*, 399–410.
Dickson, G. W., DeSanctis, G., & McBride, D. J. (1986). Understanding the effectiveness of computer graphics for decision support: A cumulative experimental approach. *Communications of the ACM, 29*, 40–47.
Fischoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge, UK: Cambridge University press.
Frederickson, B. L. (2000). Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition and Emotion, 14*, 577–606.
Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology, 65*, 45–55.
Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704.
Hammond, K. R. (1971). Computer graphics as an aid to learning. *Science, 172*, 903–908.
Jansen, S. J. T., Kievit, J., Nooij, M. A., & Stiggelbout, A. M. (2001). Stability of patients' preferences for chemotherapy: The impact of experience. *Medical Decision Making, 21*, 295–306.
Jarvenpaa, S. L., & Dickson, G. W. (1988). Graphics and managerial decision making: Research based guidelines. *Communications of the ACM, 31*, 764–774.
Kahneman, D. (1999). Objective happiness. In E. Diener, D. Kahneman, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 3–25). New York: Russell Sage.
Kahneman, D. (2000a). Experienced utility and objective happiness: A moment-based approach. In D. Kahneman & A. Tversky (Eds.), *Choices, values, and frames* (pp. 673–692). New York: Cambridge University Press.
Kahneman, D. (2000b). Evaluation by moments: Past and future. In D. Kahneman & A. Tversky (Eds.), *Choices, values, and frames* (pp. 693–708). New York: Cambridge University Press.
Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*, 697–720.
Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge: Cambridge University Press.
Kahneman, D., Frederickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4*, 401–405.
Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics, 112*, 375–405.
Kleimuntz, D. N., & Schkade, D. A. (1993). Information displays and decision processes. *Psychological Science, 4*, 221–227.
Martignon, L. (2001). Comparing fast and frugal heuristics and optimal models. In R. Selten & G. Gigerenzer (Eds.), *Bounded rationality: The adaptive toolbox* (pp. 147–171). Cambridge: MIT Press.
Morewedge, C.K., Kassam, K.S., Hsee, C.K., & Caruso, E.M. (in press). Duration sensitivity depends on stimulus familiarity. *Journal of Experimental Psychology: General*.
Pashler, H. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.
Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York, NY: Cambridge University Press.

[7] Note that this example was either in numerical form (presented here), or graphical form (see Fig. 1).

[8] A portion of this Appendix can be found in the *Journal of Behavioral Decision Making*, Experiences extended across time: Evaluation of moments and episodes, Carol Varey & Daniel Kahneman, Copyright © 1992 John Wiley & Sons Limited. Reproduced with permission.

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

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

Rickard, T. C. (2004). Strategy execution in cognitive skill learning: An item-level test of candidate models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 65–82.

Russo, J. E. (1977). The value of unit price information. *Journal of Marketing Research, 14*, 193–201.

Russo, J. E., & Leclerc, F. (1991). Characteristics of successful product information programs. *Journal of Social Issues, 47*, 73–92.

Schreiber, C. A., & Kahneman, D. (2000). Determinants of the remembered utility of aversive sounds. *Journal of Experimental Psychology: General, 129*, 27–42.

Schkade, D. A., & Kleimuntz, D. N. (1994). Information displays and choice processes: Differential effects of organizations, form, and sequence. *Organizational Behavior & Human Decision Processes, 57*, 319–337.

Simkin, D., & Hastie, R. (1987). An information-processing analysis of graph perception. *Journal of the American Statistical Association, 82*, 454–465.

Sloman, S. A. (2002). Two systems of reasoning. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 379–398). Cambridge: Cambridge University Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645–726.

Stone, A. A., Broderick, J. E., Kaell, A. T., DelesPaul, P. A. E. G., & Porter, L. E. (2000). Does the peak-end phenomenon observed in laboratory pain studies apply to real-world pain in rheumatoid arthritics? *The Journal of Pain, 1*, 212–217.

Stone, E. R., Yates, F. J., & Parker, A. M. (1997). Effects of numerical and graphical displays on professed risk taking behavior. *Journal of Experimental Psychology: Applied, 3*, 243–256.

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