No peak-end rule for simple positive experiences observed in children and adults

Eric Y. Mah[a,1]

Daniel M. Bernstein[a]

Kwantlen Polytechnic University

Mah, E.Y., & Bernstein, D.M. (in press). No peak-end rule for simple positive experiences observed in children and adults. *Journal of Applied Research in Memory & Cognition*.

Contact Information

Corresponding Author: Daniel M. Bernstein, Email: dbernste@kpu.ca, Phone number: 604-599-3372. [a]Kwantlen Polytechnic University, 12666 72nd Avenue, Surrey, B.C. V3W 2M8.

[1] Present address: University of Victoria, 3800 Finnerty Road, Victoria, B.C. V8P 5C2.

Abstract

We investigated children and adults' tendency to rely on the most intense and final moments when judging positive experiences, a heuristic known as the *peak-end rule*. This rule allows us to judge experiences quickly, but it can bias judgments. In three experiments involving various age groups (N = 988, ages 2-89), we attempted to replicate prior findings of a peak-end rule for small and simple positive experiences—i.e., receiving small gifts (Do, Rupert, & Wolford, 2008). Based on the original study and peak-end rule predictions, we hypothesized that individuals of all ages would be less satisfied with a highly-desirable gift followed by a less-desirable gift than with a highly-desirable gift alone. We failed to observe the peak-end rule in preschoolers, school-aged children, and younger/older adults in any of the contexts we investigated. Our results show little support for positive peak-end rule effects and mark possible boundary conditions for the rule.

Keywords: peak-end rule, children, adults, cognition, heuristic

General Audience Summary

This research examined peoples' tendency to judge events mainly by the best/worst moment (the peak) and the last moment (the end)—a tendency known as the *peak-end rule*. The peak-end rule is generally useful for judging events quickly, but it can result in irrational judgments. Many studies have examined the peak-end rule for negative (e.g., painful or unpleasant) events, but there is relatively little research on how the peak-end rule affects positive experiences. A previous study found evidence of peak-end rule effects in children who experienced a simple positive event: receiving candy (Do, Rupert, & Wolford, 2008). Children who received a highly-desirable candy (high peak) were more satisfied than children who received that same highly-desirable candy followed by a less-desirable one (high peak and low end). This is irrational, because two candies should be more satisfying than one. We attempted to replicate and extended this work by testing the peak-end rule in various age groups in three experiments—we examined how the peak-end rule affects judgments of a small gift (toy, candy) in individuals of various ages. People either received 1) a highly-desirable gift alone, 2) a less-desirable gift alone, 3) a highly-desirable gift followed by a less-desirable one, or 4) a less-desirable gift followed by a highly-desirable one, and then rated their satisfaction with the gift. Unlike the original study, we found little evidence for the peak-end rule in children or adults. This finding suggests that the peak-end rule likely either does not apply to or has little effect on our judgments of small, simple positive experiences.

No peak-end rule for simple positive experiences in children and adults

Imagine you've just finished a meal consisting of an excellent grilled steak followed by a decent side dish. The highlight of the meal was the steak, followed by the slightly less satisfying side. After the meal, you reflect on your experience. How happy were you with your meal? Might your answer differ if you had finished the meal after the steak?

When we retrospectively judge affective experiences, research suggests that we may focus on the most affectively intense moment (the peak) and most recent moment (the end)—an effect known as the *peak-end rule* (Kahneman, 2000; 2011). In our hypothetical example, the peak-end rule suggests that you might be happier with your meal if you finished with the most pleasurable part (peak = steak) rather than with a slightly less pleasurable part (end = side). This rule was first demonstrated for unpleasant experiences in an experiment where participants preferred a longer painful ice-water experience to a shorter one because the longer experience was manipulated to end on a less painful note.(Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993).

The tendency to judge negative experiences by a retrospective 'snapshot' of specific memorable moments instead of the length of the event and overall amount of unpleasantness experienced is counterintuitive (Kahneman, 2000). The peak-end rule affects judgments in different negative contexts, including videos, medical procedures, noises, advertisements, and learning experiences (Finn, 2010; Kahneman, 2011; Hoogerheide & Paas, 2012). Although the peak-end rule often biases judgments, it can be a useful heuristic. Accurately reconstructing an affective event and performing a hedonic calculation based on total duration and the quality of each individual moment is slow and inefficient, if not impossible. Judging an event based on the quality of the highly-memorable peak and end moments is fast and efficient (Kahneman, 2000).

However, these savings come with costs: the heuristic occasionally results in judgments that might seem irrational. A central aspect of the way we remember events is duration neglect: the duration of affective events is often heavily discounted such that longer painful episodes can be judged as more pleasant than shorter ones if the final moments of the longer episode are less painful (Kahneman, 2000).

Although the peak-end rule for negative experiences has been studied extensively, the peak-end rule for positive experiences has been relatively neglected. Some studies have found evidence of a positive peak-end rule. For example, consider the "James Dean effect": people rated a fictitious life that ended sooner but on a highly positive note as more pleasurable and desirable than the same life that ended after five additional pleasurable, but less happy years (Diener, Wirtz, & Oishi, 2001). The peak-end rule has also been observed for pleasant learning experiences (Hoogerheide & Paas, 2012), positive peer assessments (Hoogerheide, Vink, Finn, Raes, & Paas, 2017), and memories of musical pieces (Rozin, Rozin, & Goldberg, 2004). Conversely, other work has found little to no peak-end rule effects on vacation memories or judgments of pleasurable meals (Kemp, Burt, & Furneaux, 2008; Rode, Rozin, & Durlach, 2007)—eating a so-so side dish after a fantastic steak might not spoil your memories of the meal after all.

One study examined how the peak-end rule affects judgments of gifts. In one experiment, undergraduates received one of various combinations of highly-rated and lower-rated DVDs (of their choice); in another experiment, child trick-or-treaters received one of various combinations of a highly-desirable candy (Hershey's chocolate bar) and/or a less-desirable candy (bubblegum). Consistent with the peak-end rule, young adults who received a great DVD followed by an average DVD of their choice were less satisfied than those who

received just one great DVD. Similarly, children who received a Hershey's bar followed immediately by bubblegum were less satisfied than children who received the Hershey's bar alone (Do, Rupert, & Wolford, 2008).

These experiments are of interest because they appear to show that the peak-end rule affects a class of events qualitatively different from peak-end events that have typically been studied. Do and colleagues' (2008) experiments involve discretely segmented events that are shorter and simpler than typical peak-end experiences. Most peak-end experiments involve comparatively longer events that are either continuous (e.g., cold-water task in Kahneman et al., 1993) or that consist of a series of discrete events (e.g., sequences of positive or negative peer ratings in Hoogerheide et al., 2017). The DVD and candy events of Do and colleagues (2008) only involve a series of events in the loosest sense and arguably lack a meaningful duration (to be neglected). A further disconnect from prior peak-end work is that the DVD and candy experiments did not involve "direct" experiences of the stimuli (like in the original experiment). Finally, at least in the candy experiment, there appears to have been no delay between the receipt and rating of the candy. Most peak-end experiments involve a delay to make the evaluation retrospective and set the stage for heuristic remembering.

Nonetheless, Do and colleagues (2008) observed what resembles a peak-end rule in immediate evaluations of short, simple and discrete events—exactly the type of situation in which one would *not* expect to find a memory heuristic. If the effects that Do and colleagues (2008) observed are robust and replicable, there are important implications for how we judge simple positive experiences. Firstly, Do and colleagues' (2008) work suggests that the peak-end rule can distort evaluations of very simple and short discrete events that are not "directly" experienced. On a more theoretical level, by utilizing a simple event with negligible duration,

their findings show the power of salient moments, largely separate from the duration neglect aspect of the peak-end rule. In Do and colleagues' (2008) peak-end experiment, some participants experienced a longer event with a less pleasurable end. However, Do and colleagues' focus was more on the manipulated quality of the peak and end moments than the added duration. Despite the potential implications, their experiments were underpowered. Assuming a medium-sized effect (d = .5) and a one-tailed test of the critical peak-end difference, Do and colleagues' DVD experiment attained a power of only ~.53, and their candy experiment reached a power of only ~.3. If we assume a large-sized effect (d = .8), the DVD experiment has acceptable power (~.87), but the candy experiment remains woefully underpowered (~.55)[2]. Thus, more powerful tests of Do and colleagues' (2008) predictions are necessary.

Our primary aim was to provide a more comprehensive and adequately-powered test of the peak-end rule predictions examined by Do and colleagues (2008). In doing so, we extended the original paradigm to include different age groups, event timing, and stimuli. Based on Do and colleagues' (2008) findings and other prior research, we predicted that people of all ages would demonstrate the peak-end rule for small, simple positive experiences (i.e., people who receive a single, highly-desirable gift would be more satisfied than those that received a highly-desirable gift followed by a less-desirable one) . We tested this prediction in three experiments.

EXPERIMENT 1

Method

Participants

We determined sample sizes for our experiments using *a priori* power analyses via G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007). In the original experiment, Do and

[2]Power analyses conducted via G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007), as if they were *a priori* (i.e., power was adjusted under different effect sizes until the *N*'s of the original study were reached.

colleagues (2008) observed a very large peak-end effect (d = 1.35, f = .68). In light of the uncertainty surrounding that effect and possible effect size inflation due to low power, we assumed a more conservative "medium" effect size for power analyses in the current study (f = .25).

In Experiment 1, we tested 608 individuals ages 2 to 97 years drawn from a community and university population. After excluding participants with missing data on any peak-end variables except preference (n = 60), and those who indicated a preference for both or neither of the candies, n = 7), our final sample included 560 individuals. We divided this sample into four age groups: 2-7 year olds (n = 119, age m = 5.63, SD = 1.39, 50% female), 8-15 year olds (n = 128, age m = 11.51, SD = 2.29, 44% female), 16-50 year olds (n = 156, age m = 33.20, SD = 11.33, 74% female), and 51+ year olds (n = 157, age m = 67.19, SD = 9.19, 71% female). We chose these age groups to permit rough comparisons to age groups tested in Experiments 2 and 3. The total sample size exceeded the required N of 260 to achieve a power of .8 to detect all potential main and interaction effects in a 4 (Candy combination) X 4 (Age group) between-subjects ANOVA. We pre-registered Experiment 1 midway through data collection to reflect design changes and all subsequent experiments via the Open Science Framework. The pre-registration and data for our pilot test and Experiments 1-2 can be viewed here: https://osf.io/avpvx/. The pre-registration and data for Experiment 3 can be viewed here: https://osf.io/smp7d/. See Supplementary Material A for more details about our pre-registrations.

Materials & Procedure

Experiment 1 examined the peak-end rule (as it applies to receiving candy) in children and adults. Our materials and procedure for the main experiment closely followed those of the

original experiment (Do et al., 2008: Experiment 2) with minor changes to the stimuli (i.e., different candies). We used Mars bars instead of Hershey's to avoid potential nut allergies, and substituted lollipops for bubblegum because some people (e.g., younger children, older adults) may have had trouble chewing the gum. We made these stimulus changes for all subjects in the interests of safety (e.g., researcher error or participant failure to report allergies). We validated these stimuli (i.e., that they were viewed as highly and less desirable, respectively) via pre-registered pilot testing (see Supplementary Material B for more details). For the main experiment, participants were assigned to one of four Candy combinations: "A" the "high peak" (a highly-desirable Mars bar), "B" the "low peak" (a lollipop), "A_B" the "high peak-low end" (a highly-desirable Mars bar followed by a less-desirable lollipop), or "B_A" the "high peak-high end" (lollipop followed by Mars bar).

For all Candy combinations, we told participants that they would receive a small prize for participating in the longer experiment, and then gave them the Mars bar (or lollipop). For combinations A_B and B_A, we remarked that we had an additional candy for the participant ("I have another candy for you"), and we then gave participants the lollipop (or Mars bar). After receiving the candy, participants rated how "they felt about the candy [candies] we gave them." Participants indicated their rating on the same 7-point smiley face scale used in Do and colleagues' experiment (2008), which ranged from "Neither happy nor unhappy" to "Very happy". The satisfaction question was worded in a way to encourage participants to rate the single candy on its own and the two candies together (i.e., in the single-candy conditions, participants rated their satisfaction with the "candy" we gave them; in the dual-candy conditions, participants rated their satisfaction with both "candies" we gave them). Though we did not include a preference question initially in the experiment, we introduced such a question partway

through data collection. For this question, after rating the gift, participants indicated which of the two candies they preferred (participants given a single candy at this point received the candy they did not originally rate). This question was aimed at supplementing our pilot data on candy preference rates and was used to exclude participants who indicated a preference for the lollipop over the Mars bar (or a preference for both or neither candy). Because we introduced this preference question partway through the study, we only obtained preference data for roughly 40% (n = 234) of our total sample. Participants completed our candy peak-end task at the end of a longer 60-90 minute cognitive psychology experiment.

Note that in both the original experiment and our replication, participants did not eat the candy as part of the experiment or rating procedure. As such, the positive event we examined was the experience of receiving a gift, not eating candy. However, prior studies have shown that directly experiencing an event may not be necessary for the peak-end rule to occur (e.g., hypothetical meals; Rode et al., 2007, imagined lives; Diener et al., 2001)

Results

We predicted that satisfaction ratings of the highly-desirable Mars bar ("A") would be higher than ratings of the less-desirable lollipop ("B") *and* higher than ratings of the Mars bar followed by the lollipop ("A_B"). Recall that the peak-end rule predicts that people mainly focus on the peak (best moment) and the end (final moment) when making overall retrospective judgments about affective experiences. Accordingly, ending with the lollipop should reduce satisfaction ratings. However, beginning with the lollipop should not lower ratings, because the highly-desirable Mars bar serves as both the peak and the end. Thus, we also predicted that those who received the Mars bar after the lollipop would be more satisfied than those who received the lollipop after the Mars bar. Finally, we predicted that the peak-end rule would

operate similarly across all age groups (i.e., no interaction between age group and Candy combination).

We first examined candy preference data. Though we only had preference data for about half of our participants, those who answered the preference question preferred the Mars bar (83% of 8-15 year olds, 79% of 16-50 year olds, and 93% of 51+ year olds). The exception was 2-7 year olds, who preferred the Mars bar (56%) only slightly more than the lollipop. See Supplementary Material 1A for all percentages by age group. In total, 48 participants indicated a preference for the lollipop. Though the original pre-registered plan was to perform separate analyses on participants who indicated a preference for the lollipop, we did not collect enough data from "lollipop-preferers" to warrant a separate analysis. Instead, we reversed their conditions (e.g., we treated the lollipop as highly-desirable and the Mars bar as less-desirable for these participants) and combining them with participants who indicated a preference for the Mars bar. A cursory manipulation check in our lollipop-preferring subsample suggested that a simple condition reversal was valid. Those in the highly-desirable "A" condition (lollipop for these participants) gave significantly higher satisfaction ratings than those in the less-desirable "B" condition (Mars bar for these participants), $t(22.97) = 2.21$, $p = .04$. Though a corresponding Wilcoxon rank-sum test (in light of dubious normality and variance homogeneity) was *not* statistically significant ($p = .07$), the difference was in the appropriate direction.

We analyzed Candy combination satisfaction ratings by age group (4 X 4 between-subjects ANOVA), and found a significant main effect of age group, $F(3, 544) = 11.90$, $p < .001$, partial $\eta^2 = .05$. Due to evidence of assumption violations (heterogeneity of variance and non-normality in satisfaction ratings across Candy combinations and age groups), we opted to use non-parametric Kruskal-Wallis tests to follow up on significant main effects/interactions. In all

cases, parametric and non-parametric results agreed (results listed in Supplementary Material 1B). We also used Wilcoxon rank-sum tests in place of planned and unplanned pairwise comparisons (with Bonferroni-corrected alpha levels). Following up on our age group main effect: 2-7 year olds were more satisfied with candy than 16+ year olds ($p \leq .001$), but there were no other age group differences. There was also a main effect of Candy combination, $F(3, 544) = 9.81$, $p < .001$, partial $\eta^2 = .05$. We successfully manipulated candy desirability (A > B, $p < .001$, Cohen's $d = .71$ (95% CI: .39, 1.03)) but observed no peak-end differences. There was no difference between A and A_B, $p = .994$, Cohen's $d = -0.04$ (95% CI: -.23, .15) or between A_B and B_A, $p = .434$, Cohen's $d = -.19$ (95% CI: -.52, .12)). Finally, there was no significant interaction between age group and Candy combination, $F(9, 544) = 1.24$, $p = .27$. We also conducted the above analyses with only participants who indicated a preference for the Mars bar and obtained identical results. Table 1 lists overall Candy Combination means collapsing across age groups (with the critical conditions bolded).

Table 1. *Mean satisfaction ratings by Candy combination and age group*

Candy Combination	Mean	SD	*n*
A	**5.72**	**1.54**	**220**
B	4.53	2.02	64
A_B	**5.76**	**1.41**	**217**
B_A	5.59	1.55	59

Note. A = Highly-desirable candy, B = Less-desirable candy, A_B = Highly-desirable candy followed by less-desirable candy, B_A = Less-desirable candy followed by highly-desirable candy.

Because we lacked preference data for most of our sample, it is likely that there were undetected lollipop-preferers for whom we couldn't appropriately recode conditions. Based on our cursory analysis of the Mars bar vs. lollipop difference in these participants, it is likely that lollipop preferers' response pattern was opposite that of the full sample. This could have plausibly led to the attenuation of group differences. However, recall that the vast majority of participants who *were* asked about their preference indicated that they preferred the Mars bar (≥ 83% in all but the youngest age group). We also found a consistent A > B difference in the overall sample, and our pilot testing suggests that our stimulus manipulation worked as intended. Given all this, we think it unlikely that undetected "lollipop-preferers" in our full sample are the culprits behind our null results.

The standard null hypothesis significance testing (NHST) approach does not allow one to draw substantive conclusions when null results emerge—*p* values only quantify evidence *against* the null and cannot be used to corroborate it (Loftus, 1996). Bayesian hypothesis testing is a compelling alternative to the standard NHST approach and offers a way of quantifying relative evidentiary support for competing hypotheses—for example, H_0 vs. H_1. Aside from allowing one to directly compare the plausibility of competing hypotheses, Bayesian inference offers several advantages over NHST (for a comprehensive review see Wagenmakers et al., 2018).

To better understand our null results, we supplemented our NHST analyses with Bayesian tests. For our first Bayesian analysis, we tested the following hypotheses: H_0 (that there is no peak-end rule for small positive experiences), and H_1 (that there is a peak-end rule for small positive experiences). The resulting *Bayes factor* (BF) for this analysis quantifies the relative support for one hypothesis over another. We obtained a BF_{01} ($_{01}$ indicating support in favor of H_0 relative to H_1) of 12, indicating that given our data, we should update our belief in no

peak-end rule relative to our belief in a peak-end rule by a factor of 12:1. This constitutes strong evidence for a null effect (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Details about this analysis (e.g., prior choice) and Bayesian analyses in general (e.g., implications of priors) can be found in our Supplementary Material (1A).

In addition to a standard Bayesian analysis, we also examined how our results differed across a wide range of priors (i.e., pre-data assumptions about the peak-end rule effect size). For most reasonable prior choices, this "robust" Bayesian analysis resulted in moderate-to-strong evidence for H_0 (see Supplementary Material 1B for more details about this analysis). Finally, we conducted two additional analyses using non-standard priors that represented more specific predictions one might have about the peak-end rule, based specifically on the effect size observed in Do and colleagues (2008). The resulting BF_{01}s for these analyses were 7.69 and 5—in both cases, moderate evidence for H_0. The details for these analyses, including the priors used and background for non-standard prior analyses, can be found in the Supplementary Material (1C).

Though the results of our omnibus Candy combination X age group test pertain less directly to the peak-end rule, our decision to collapse across age groups relies on the assumption of no interaction. Using a Bayesian ANOVA with default specifications (see Rouder, Morey, Verhage, Swagman, & Wagenmakers, 2016 for an overview), we obtained a BF_{01} of 19.94. In other words, a model of the peak-end rule *without* a Candy combination X age interaction is 19.94 times as likely as a model with such an interaction. This analysis provides strong evidence that the peak-end rule (lack thereof) does not differ across our age groups

Discussion

The peak-end rule that Do and colleagues (2008) observed failed to materialize in our considerably larger, all-ages sample. Our NHST analyses suggest the absence of, or a trivially small, peak-end rule effect. A variety of Bayesian analyses provided moderate-to-strong evidence *against* a peak-end rule across all age groups. Even though we successfully manipulated candy desirability (as evidenced by our A > B difference and our pilot testing), the addition of the less-desirable candy to the highly-desirable one did not produce a less satisfying experience.

EXPERIMENT 2

We did not observe the peak-end rule for small positive experiences in any of the age groups we tested. However, in Experiment 1, the peak-end task followed a 60-90 minute study. Remembering events is prerequisite to judging them. Consequently, memory for events depends on the ability to segment experiences into discrete events (Sargent et al., 2013). Thus, the candy event could have been judged as part of the larger psychology study experience. Experiment 2 addressed this possibility by comparing Candy combinations either at the beginning or end of a longer experience. For Experiment 2, our predictions were the same as Experiment 1 (i.e., A > B, A > A_B, A = B_A), but only when the peak-end task occurred before a longer experience. When the peak-end task occurred after a long experience we predicted that people would not show peak-end effects—only the non-peak-end differences we observed in Experiment 1.

Method

Participants

Participants in Experiment 2 included 277 undergraduate students. After excluding 8 participants who indicated a preference for both or neither candy, our sample included 269 participants. Of these participants, 66 (24.5%) indicated a preference for the lollipop. Like

Experiment 1, our original analysis plan was to analyze lollipop-preferers as a separate sample. Due to the low *n* in this subsample, we instead considered reversing their conditions (i.e., the lollipop is treated as the highly-desirable candy for these participants, and the Mars bar the less-desirable candy). However, an exploratory manipulation check in this subsample was not successful, $t(24.63) = .06$, $p = .95$ (the corresponding Wilcoxon rank-sum test was also non-significant, $p = .98$). On this basis, we could not justify a simple condition reversal for lollipop-preferers in our sample. Thus, we chose to exclude these participants from subsequent analyses. Our sample size of 203 (Age $M = 21.29$, $SD = 4.25$, 81% female) exceeded the recommended *N* of 179 to achieve a power of .8 to detect all potential main and interaction effects for a 4 (Candy combination) X 2 (Timing) between-subjects ANOVA.

Materials and Procedure

Student participants completed the same peak-end task used in Experiment 1 as part of 1 of 4 longer psychology experiments that they completed for course credit (average experiment length = ~40 minutes). Experiment 2's procedure was identical to Experiment 1's except that participants received one of the four candy combinations either at the beginning or end of one of the psychology experiments. As participants received the candy/candies, the experimenter mentioned that the candy was a thank-you gift for participating in the experiment.

Results

For our main hypothesis, we predicted that the peak-end rule is reduced or eliminated if the peak-end experience occurs after another longer experience. To test this, we used a 4 (Candy combination) X 2 (Timing) between-subjects ANOVA. We found a main effect of Candy combination, $F(3, 195) = 4.82$, $p = .003$, partial $\eta^2 = .07$, but no interaction between Candy combination and timing, $F = .57$, $p = .638$. A Shapiro-Wilk test suggested non-normality in our

data, but because cell sizes were essentially equal and homogeneity of variance was present, we used only parametric tests for Experiment 2 analyses. Planned follow-up *t*-tests revealed a successful candy desirability manipulation (A > B, $t(100) = 3.55$, $p < .001$(Cohen's $d = .70$, 95% CI: .30, 1.11). However, we did not observe a peak-end rule; there was no significant difference between A and A_B, $t(100) = 0.48$, $p = .63$ (Cohen's $d = .09$, 95% CI: -.30, .49); or between A_B and B_A, , $t(99) = -.80$, $p = .43$ (Cohen's $d = .16$, 95% CI: -.24, .55). We did find a significant main effect of timing, $F(1, 195) = 7.81$, $p = .006$, partial $\eta^2 = .04$, such that satisfaction with the candy gift was higher at the end of a longer experience than at the beginning. Table 2 lists Candy combination mean satisfaction ratings, overall and by timing.

Table 2. *Mean Satisfaction ratings by Candy combination and Timing*

Candy combination	**Overall (SD) [*n*]**	**Beginning of Study (SD) [*n*]**	**End of Study (SD) [*n*]**
A	**6.02 (1.20) [52]**	5.93 (1.36) [30]	6.14 (.94) [22]
B	5.08 (1.47) [50]	4.88 (1.48) [24]	5.27 (1.46) [26]
A_B	**5.90 (1.33) [50]**	5.50 (1.42) [26]	6.33 (1.09) [24]
B_A	5.68 (1.48) [51]	5.32 (1.57) [25]	6.02 (1.32) [26]
Overall	*5.67 (1.41) [203]*	*5.44 (1.49) [105]*	*5.92 (1.28) [98]*

Note. A = Mars Bar, B = Lollipop, A_B = Mars then Lollipop, B_A = Lollipop then Mars

As with Experiment 1, we supplemented our NHST analysis with Bayesian analyses. First, it was crucial to determine whether we could rule out the possibility of timing effects on the peak-end rule (i.e., a Candy combination X timing interaction). A Bayesian ANOVA analogous to the one conducted in Experiment 1 returned a BF_{01} of 8.03—moderate evidence against such an interaction. Next we examined the lack of observed differences between our critical peak-end conditions (A vs. A_B). Using the same prior as Experiment 1, we obtained a BF_{01} of 3.21, indicating moderate evidence for H_0 relative to H_1. A robustness analysis revealed

that under more conservative prior widths the evidence in favor of H_0 was anecdotal—less strong evidence than that observed in Experiment 1 (see Supplementary Material 2A for the robustness plot).

Across a wide range of reasonable prior widths, we failed to observe more-than-anecdotal evidence against the peak-end rule—weaker than the evidence observed in Experiment 1. Observing compelling evidence for null or trivially small effects with Bayesian analyses typically requires larger sample sizes to detect small deviations from the null (e.g., in excess of N = 50,000 to detect effects of d = .02 or smaller; Rouder et al., 2009). Thus, it is likely that our results are due to Experiment 2's smaller sample size (rather than any substantive differences in peak-end rule effects across experiments).

To see how our Experiment 2 results should affect our post-Experiment 1 beliefs in the peak-end rule, we conducted an informed Bayesian analysis. For this analysis, the prior distribution was the distribution of likely effect sizes given the Experiment 1 results[3]. Here the BF represents the change in our beliefs in a null effect before and after the Experiment 2 data. We obtained a BF_{01} of 1.01, suggesting that the Experiment 2 results tell us little beyond what we learned from Experiment 1. Again, this is unsurprising given the smaller sample size of Experiment 2. Though it might seem intuitive to characterize our current belief in the peak-end rule as a simple product of the BF_{01}s observed in both experiments (e.g., Experiment 1 belief in H_0:H_1 of 12:1 * Experiment 2 updating factor of 1.01), doing so is not valid (Rouder & Morey, 2011). Individual experiment BFs respect sample size—with smaller samples, small effects are more likely to be considered as evidence for the null (2011).

Discussion

[3] Specifically, a *t*-distribution centered on the A vs. A_B raw scale difference observed in Experiment 1, with a *SD* equal to the standard error of that difference and *df* equal to the Experiment 1 analysis *df*.

Our Experiment 2 analyses failed to provide evidence for the peak-end rule regardless of event timing. This suggests that the results of Experiment 1 were not substantially tainted by the experimental context and thus reflect a meaningful null effect.

EXPERIMENT 3

Experiment 1 provided strong evidence *against* a peak-end rule for small positive experiences, and Experiment 2 provided no support for the peak-end rule. However, one could argue that there is something about candy in general that precludes peak-end effects. Therefore, we endeavoured one final test of the peak-end rule, this time by giving small, non-candy gifts (toys) to children and youth. Our hypotheses mirrored those of Experiments 1 and 2: We predicted that children (3-7 year olds) and youth (8-15 year olds) would be less satisfied with a highly-desirable gift followed by a less-desirable gift (A_B) than a highly desirable gift alone (A). We included age group as a factor but did not predict any age group differences in the peak-end rule.

Method

Participants

Our sample included 117 children ages 3-7 and 121 children ages 8-15. Examining toy preference data, we found that 82% of 3-7s who received both toys preferred A to B (i.e., the glass rock to the brown paper bag), while 94% of 8-15s preferred A to B (i.e., the magnetic rocks to the wooden craft stick). Because we lacked enough participants to evaluate our manipulation in the subgroups with reversed preference, we excluded from further analysis all participants who indicated a preference for B (3-7s n = 10, 8-15s n = 3). Thus, our final sample included 107 children ages 3-7 (Age m = 5.08, SD = 1.41, 41.3% female) and 118 children ages 8-15 (Age m = 9.57, SD = 1.68, 44.7% female). Our sample exceeded the recommended N of 179 to attain a

power of .8 to detect all potential main and interaction effects in a 4 (Toy combination) X 2 (age group) between-subjects ANOVA. The study was run at a local science centre, where we approached parents entering the centre and asked if their children would be interested in participating in a short psychology study on how children and youth rate small toys. Specifically, we told children that we were going to give them a small gift and that we would ask them to tell us how they felt about it (children in the dual toy condition were not initially told they would receive two toys). After explaining the study purpose and obtaining consent, children received one of the toy combinations and were asked to tell us how they felt about the toy/toys. Though we had no control over what kind of experience children had before entering the science center, it seems unlikely that participants viewed our peak-end study as a continuation of a longer, prior experience—at least not in the same way participants may have viewed the peak-end experience in Experiments 1 and 2. Additionally, real-world affective experiences necessarily occur before and after other affective experiences.

Materials and Procedure

We conducted pilot testing to select toys and found that the age groups differed in terms of what toys they preferred. Our final highly-desirable toys were: glass rocks for 3-7 year olds and magnetic rocks for 8-15 year olds; our less-desirable toys were: brown paper bags for 3-7 year olds and wooden craft sticks for 8-15 year olds. Other than the stimulus change, the experimental procedure was identical to Experiment 1, except that we only collected preference data from participants who received two toys (due to costs associated with having to give all children both toys). Again, it is worth noting that children were rating their experience of receiving the gifts (i.e., as opposed to playing with and then rating gifts).

Results

To test whether children and youth showed peak-end rule effects for a small positive experience *other* than candy, we conducted a 4 (Toy combination) X 2 (Age group) between-subjects ANOVA. We found a main effect of Age group, $F(1, 217) = 9.34$, $p = .003$, partial $\eta^2 = .03$. Overall, 3-7 year olds were happier with the prizes than 8-15 year olds. We also found a main effect of Toy combination, $F(3, 217) = 8.84$, $p < .001$, partial $\eta^2 = .07$.

However, both our main effects were qualified by a significant interaction, $F(3, 217) = 6.95$, $p < .001$, partial $\eta^2 = .09$. A follow-up one-way ANOVA comparing Toy combination ratings in 2-7 year olds revealed that ratings for all Toy combinations did not significantly differ, $F(3, 103) = .355$, $p = .786$.. Youth ages 8-15 drove our Toy combination main effect, with a significant omnibus test in this age group, $F(3, 114) = 16.06$, $p < .002$. Planned follow-up tests (Bonferroni-corrected *t*-tests) revealed that we successfully manipulated toy desirability, $t(68) = 5.33$, $p < .002$ (Cohen's $d = 1.36$, 95% CI: .80, 1.91). However, despite successfully manipulating toy desirability in 8-15 year olds, we observed no peak-end rule in this age group (No difference between A and A_B, $t(43) = 1.75$, $p = .09$, Cohen's $d = .52$, 95% CI: -.09, 1.13) or between A_B and B_A ($t(46) = .11$, $p = .91$, Cohen's $d = .03$, 95% CI: -.55, .62). Table 3 lists mean Toy combination satisfaction ratings by Age group.

Table 3. *Mean satisfaction ratings by Toy combination and Age group*

Toy combination	**3-7s (SD) [*n*]**	**8-15s (SD) [*n*]**
A	5.52 (1.94) [33]	**6.04 (1.02) [23]**
B	5.76 (1.83) [29]	**3.70 (1.98) [47]**
A_B	6.00 (1.60) [23]	**5.50 (1.06) [22]**
B_A	5.64 (1.65) [22]	**5.46 (1.30) [26]**

Note. A = Highly-desirable toy (Glass rocks for 3-7s, Magnetic rocks for 8-15s), B = Less-desirable toy (Brown paper bags for 3-7s, Wooden craft sticks for 8-15s).

Though we couldn't fully investigate the peak-end rule in younger children due to unsuccessful manipulation of toy desirability, we successfully manipulated toy desirability in older children. They rated the magnetic rocks (A) as highly-desirable (M = 6.04 out of 7), and the wooden craft sticks (B) as less-desirable (M = 3.70 out of 7). This difference is not only significant, but more in line with the differences observed in Do and colleagues' (2008) original study. Despite this, we found no peak-end rule: Older children did not rate A alone as better than A_B, as the peak-end rule would predict.

Despite Experiment 3's smaller sample size, we conducted a final Bayesian test of the critical peak-end conditions (A vs. A_B). Using a default prior, we obtained a BF_{01} of 0.53, indicating anecdotal support for a peak-end rule. However, a corresponding robust version of this analysis failed to produce more-than-anecdotal evidence in favor of a peak-end rule for the entire range of prior widths we examined (See Supplementary Material 3A for the robustness plot). Given our small sample size and mounting evidence for a likely null and possibly small true effect size, these inconclusive results are not surprising.

How should these results affect our belief in the peak-end rule? An informed Bayesian analysis using the combined results of Experiments 1 & 2 as the prior resulted in a BF_{01} of .64. In other words, our belief in H_0 is slightly reduced in light of Experiment 3's results. As previously mentioned, one cannot simply combine BFs across experiments (Rouder & Morey, 2011). To better quantify our Bayesian evidence for and against the peak-end rule across all three experiments, we conducted a final Bayesian meta-analysis. Under the default prior width, the combined BF_{01} was 7.77—moderate evidence against a peak-end rule for small positive experiences (See Supplementary Material 3B for the robustness plot).

Discussion

Again, our results failed to support a peak-end rule in children for toys. Though the Experiment 3 peak-end difference (A > A_B) was the largest numerically out of all three experiments, it was not significant. Additionally, there is an unexpected difference we observed that runs counter to the peak-end rule: the anomalous A > B_A (numerical) difference. Recall that the peak-end rule predicts no difference between A & B_A (as the peak and end in both cases are the same). Given that this anti-peak-end rule numerical difference is of similar magnitude to the peak-end rule difference, and Experiment 3's smaller sample size, we are cautious to conclude that Experiment 3 provides more support for the peak-end rule than the other experiments. Finally, though the individual BF test of Experiment 3 was uninformative (likely due to small sample size), the sum of our Bayesian evidence weighed against a peak-end rule for small positive experiences.

General Discussion

We attempted to replicate and extend findings of a peak-end rule for small positive experiences and failed to find any effects in children or adults (or any of the various age groups examined in Experiment 1) with two different types of small gifts (candies and toys). People of all ages who received a highly-desirable gift were no more or less satisfied than people who received the same gift followed by a less-desirable one. Our tests of candy (Experiments 1 & 2) resulted in either evidence against or a lack of evidence for the peak-end rule. Similarly, our toy experiment failed to provide convincing evidence for the peak-end rule.

Though the Bayesian evidence obtained in Experiments 2 & 3 was ambiguous, taken together, our experiments provided moderate evidence that the peak-end rule does not affect the kinds of experiences we examined.

Ultimately, we think that the most likely explanation for our failure to replicate Do and colleagues' (2008) original results is that the peak-end rule either does not apply to these kinds of experiences or has negligible effects on them. The experiences under examination were simple and brief. Other experiments have found evidence for positive peak-end effects using longer events (e.g., pleasant study sessions, DVD gifts separated by a delay; Hoogerheide & Paas, 2012; Do et al., 2008, Experiment 1). Thus, it is possible that short and simple events such as receiving gifts do not invoke the *evaluation by moments* heuristic. The qualitative difference between the gift event we studied and the affective events typically studied in the peak-end literature may explain our lack of peak-end findings. People may be able to accurately judge the affective quality of simple and discretely segmented events without relying on the peak-end rule.

One possible explanation for our null results lies with our manipulation. Because retrospective peak-end judgments are a combination of peak and end affect, ratings of an event with a similar peak and end will be like ratings of the peak or end alone (i.e., if A = B, then we would also expect A = A_B). Our manipulation of affective quality was weaker than the manipulation in the original study (Do et al., 2008). However, we still observed a significant difference in ratings between the stimuli, and the peak-end differences we observed fell far short of the magnitude one might predict (i.e., the average of A & B ratings). Additionally, the peak-end rule was originally observed with a small manipulation of stimulus quality (e.g., a gradual one-degree shift in ice water temperature; Kahneman et al., 1993). Thus, we do not think that our weaker manipulation explains our findings.

Finally, one may argue there are other paradigms more relevant to the evaluation of short, simple experiences. As we have discussed, the candy/toy events differ from typical events in the peak-end paradigm in terms of duration, continuity, complexity, and delay between event and

evaluation. It may be more appropriate to consider Do and colleagues' (2008) findings as reflecting some other process. For instance, people were less satisfied with a DVD or candy gift when given a less-desirable gift after a highly-desirable one. This finding is compatible with a general preference for "happy endings" (Ross & Simonson, 1991)—where given sequences consisting of discrete, essentially duration-less positive and negative experiences, people prefer sequences that end with a positive experience. Yet another possibility is that aversion to affectively-decreasing sequences can be explained in terms of adaptation, where affect is tuned to an initial set point by the first stimulus and even small changes in stimulus quality result in substantial changes in affect (Haisley & Loewenstein, 2011). From an adaptation perspective, people in Do and colleagues' (2008) experiment may have been initially happy with the great Hershey's bar but affect dropped when the stimulus quality of the additional lollipop departed from the higher set point. Similarly, the "peak-end" pattern observed by Do and colleagues may have been due to expectation violation. When people receive a gift and learn that another one is forthcoming, expectations about the gift-to-be are based on the gift that was already received (Haisley & Loewenstein, 2011). Thus, when expectations for another Hershey's-quality gift are subverted, satisfaction with the overall gift declines.

These alternative ideas present attractive and perhaps more appropriate ways to conceptualize evaluations of short, simple and discrete affective experiences. However, we chose to remain within the peak-end perspective adopted by Do and colleagues (2008). The primary goal of our three experiments was to investigate a surprising (and important) finding of a peak-end rule for experiences far different than those typically found in the peak-end literature. While it may seem evident *a priori* that the peak-end rule should not apply to simple positive experiences, Do and colleagues' findings challenge that reasoning. Without further

comprehensive tests supporting or opposing their results, there is a tension between *a priori* notions about the peak-end rule and empirical results. Our research serves as a stronger test of Do and colleagues' findings (i.e., larger samples, lack of rating ceiling effects, more variability in ratings), and provides evidence that the peak-end rule does not affect short positive experiences.

That is not to say that these alternative theories are irrelevant to the peak-end rule or our experiments. Though we did not test these alternate theories we believe that our results suggest that alternate biases or heuristics were not at play in the experiences we examined. Across our three experiments, participants generally viewed the highly-desirable gift as more satisfying than the less-desirable gift, but did not appear to show order effects, preference for a better end, or aversion to a worse end. However, regardless of the status of these theories relative to our results, we believe that our adequately-powered failure to replicate prior findings provides good evidence that the peak-end rule does not apply to such experiences.

Some open questions remain. It is possible that short, simple positive experiences elicit a peak-end rule, but only when a delay separates experience and evaluation. Given that the peak-end rule is based on *retrospective* evaluations, it is possible that the lack of delay in our study eliminated a peak-end rule that would otherwise exist. Because our main aim was to replicate Do and colleagues (2008), our lack of a delay manipulation leaves us unable to rule out that possibility. Outside of direct replication, the fact that Do and colleagues observed an apparent peak-end rule with no delay or negligible delay makes their finding even more surprising, and worthy of further tests. However, addressing the possibility that small positive experiences elicit a peak-end rule with longer delays is certainly a worthwhile next step in further tests of peak-end rule boundary conditions. Finally, we did not directly examine evaluations of small positive

experiences through the lenses of the alternative paradigms that we described prior. Though our results suggest that small positive experiences are relatively resilient to evaluation biases, more direct tests are necessary to make conclusive claims.

Our study provides evidence that the peak-end rule does not substantially affect our judgments of short, simple positive experiences. In failing to replicate the surprising findings of a previous peak-end study (Do et al., 2008), we highlight potential boundary conditions for the peak-end rule. Though some important questions remain, it seems likely that our in-the-moment and soon-after-the-fact judgments of these experiences are unbiased. Further research along these lines will allow us to better understand how we think about the positive experiences that shape our lives. Such research will help us decide whether we really should save the best—whether it's steak, a good movie, or a big gift—for last!

References

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

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*(781). doi: 10.3389/fpsyg.2014.00781

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review, 25*(1), 207-218. DOI: 10.3758/s13423-017-1266-z

Do, A.M., Rupert, A.V., & Wolford, F. (2008). Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic Bulletin & Review, 15*(1), 96-98. doi: 10.3758/PBR.15.1.96

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. doi: 10.3758/BF03193146

Finn, B. (2010). Ending on a high note: Adding a better end to effortful study. *Journal of Experiment Psychology: Learning, Memory, and Cognition, 36*(6), 1548-1553. doi: 10.1037/a0020605

Grange, G. (2015, November 27). Animating Robustness-Check of Bayes Factor Priors [Blog post]. Retrieved from: https://jimgrange.wordpress.com/2015/11/27/animating-robustness-check-of-bayes-factor-priors/

Haisley, E., & Loewenstein, G. (2011). It's not what you get but when you get it: The effect of gift sequence on deposit balances and customer sentiment in a commercial bank. *Journal of Marketing Research, 48*(1), 103-115. https://doi.org/10.1509/jmkr.48.1.103

Hoogerheide, V., & Paas, F. (2012). Remembered utility of unpleasant and pleasant learning experiences: Is all well that ends well? *Applied Cognitive Psychology, 26*(6), 887-894. doi: 10.1002/acp.2890

Hoogerheide, V., Vink, M., Finn, B., Raes, A.K., & Paas, F. (2017). How to bring the news … peak-end effects in children's affective responses to peer assessments of their social behavior. *Cognition & Emotion, 32*(5), 1114-1121. doi:10.1080/02699931.2017.1362375

Kahneman, D. (2000). Evaluation by moments: Past and future. In D. Kahneman & A. Tversky (Eds.), *Choices, Values and Frames* (693-708). New York: Cambridge University Press and the Russell Sage Foundation.

Kahneman, D. (2011). *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux.

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. doi: 10.1111/j.1467-9280.1993.tb00589.x

Kemp, S., Burt, C.D.B., & Furneaux, L. (2008). A test of the peak-end rule with extended autobiographical events. *Memory & Cognition, 36*(1), 132-138. doi: 10.3758/MC.36.1.132

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*(6), 161-171. https://doi.org/10.1111/1467-8721.ep11512376

Rode, E., Rozin, P., & Durlach, P. (2007). Experienced and remembered pleasure for meals: Duration neglect but minimal peak, end (recency) or primacy effects. *Appetite, 49*(1), 18-29. doi:10.1016/j.appet.2006.09.006

Ross, W. T., & Simonson, I. (1991). Evaluations of pairs of experiences: A preference for happy endings. *Journal of Behavioral Decision Making, 4(4), 273–282.* doi:10.1002/bdm.3960040405

Rouder, J.N., & Morey, R.D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18*(4), 682-689. doi: 10.3758/s13423-011-0088-7

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237. doi:10.3758/PBR.16.2.225

Rouder, J.N., Morey, R.D., Verhagen, J., Swagman, A.R., & Wagenmakers, E-J. (2016). Bayesian analysis of factorial designs. *Psychological Methods, 22*(2), 304-321. DOI: 10.1037/met0000057

Rozin, A., Rozin, P., & Goldberg, E. (2004). The feelings of music past: How listeners remember musical affect. *Music Perception: An Interdisciplinary Journal*, *22*(1), 15-39. DOI: 10.1525/mp.2004.22.1.15

Sargent, J.Q., Zacks, J.M., Hambrick, D.Z., Zacks, R.T., Kurby, C.A., Bailtey, H.R., . . . Beck, T.M. (2013). Event segmentation ability uniquely predicts event memory. *Cognition, 129*(2), 241-255. doi: 10.1016/j.cognition.2013.07.002

Schönbrodt, F. (2013, August 23). Exploring the robustness of Bayes Factors: A convenient plotting function [Blog post]. Retrieved from: https://www.nicebread.de/exploring-the-robustness-of-bayes-factors-a-convenient-plotting-function-2/

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van Der Maas, H.L.J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on

Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426-432. doi: 10.1037/a0022790

Wagenmakers, E.-J. (2015, November). Cauchy prior widths [Online Forum Post]. Message posted to https://forum.cogsci.nl/index.php?p=/discussion/1725/cauchy-prior-widths.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . ., & Rouder, J.N. (2018). Bayesian inference for psychology: Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35-57. DOI 10.3758/s13423-017-1343-3

Wiens, S. (2017, May 6). Aladins Bayes Factor in R (Version 3). figshare. https://doi.org/10.17045/sthlmuni.4981154.v3

Author Contributions and Acknowledgments

Daniel Bernstein conceived of and designed the study, collected the data and assisted with the writing of the manuscript. Eric Mah assisted with the conception and design of Experiments 2 & 3, collected, analyzed and interpreted the data, and wrote the manuscript. We acknowledge Maryanne Garry for her helpful suggestions regarding the writing and data analysis, and André Aßfalg for his helpful suggestions regarding our statistical analyses and interpretation of results. We would also like to thank TELUS World of Science for their cooperation in running Experiment 3. This work was supported by the Canada Research Chairs Program (950-228407) and the Social Sciences and Humanities Research Council of Canada (435-2015-0721).

The authors have no conflicts of interest to declare.

Supplementary Material

A. Pre-Registration

Note that we pre-registered after prior data collection and analysis for Experiment 1, so the pre-registration reflects our thinking halfway through data collection. Additionally, our final analyses deviated from the pre-registration in the following ways: 1) For Experiment 1, we initially planned to split our all-ages sample into children (ages 3-17) and adults (ages 18+). However, we were able to collect enough data to permit slightly more fine-grained age group comparisons (i.e., the four age groups included in the experiment), 2). Additionally, we omitted some measures and analyses from Experiment 3, namely, "length of Science Centre visit" and "overall satisfaction with Science Centre visit". We excluded these measures because we collected the vast majority of our data immediately as participants entered the Science Centre. Finally, we did not pre-register our Bayesian analyses. Our pre-registrations and data files can be viewed here: https://osf.io/avpvx/ (Pilot, Experiments 1 & 2) and here: https://osf.io/3fqvk/ (Experiment 3).

B. Online Pilot Testing

We conducted online pilot testing to validate our stimulus changes and ensure that the Mars bar and lollipops were likely to be viewed as "highly-desirable" and "less-desirable", respectively. The lollipops that we used also varied in terms of colour, so we needed to test whether there were systematic differences in lollipop ratings as a function of colour. We conducted this pilot study online via Amazon Mechanical Turk (mTurk) with an adult sample (N = 243, ages 19-69, age M = 35.51, SD = 10.88). mTurk is a website where people can offer and complete small "micro jobs" for pay. Online pilot participants received $0.10 for completing our brief candy questionnaire.

This pilot test exceeded the recommended sample size of $N = 200$ to achieve a power of .8 assuming a medium effect size ($g = .25$). This calculation was based on the most demanding analysis of the pilot test (comparison of 5 different lollipop colors).

B1. Methods. We used the online pilot study to test the validity of our candy stimuli. For the study, participants viewed pictures of both candies, chose whether they preferred the Mars bar or lollipop, and gave a satisfaction rating for both candies. The order of the preference and satisfaction questions were randomized. Satisfaction ratings were made on the same 7-point smiley-face scale used in Do and colleagues' (2008) candy experiment (1 = "neutral" smiley face, 7 = "very happy" smiley face). Lollipops of all 5 colours were included in the pilot (Orange, Red, Yellow, Green, Black).

B2. Results. The pilot study successfully demonstrated that participants viewed the Mars bar as more desirable than the lollipop: 81.9% of participants chose the Mars bar when offered a hypothetical choice between the two. A one sample t-test comparing our sample to a hypothetical population with neutral preferences confirmed that, given a forced choice, participants preferred the Mars bar to the lollipop, $t(242) = 12.88$, $p < .001$. A paired-samples t-test comparing continuous satisfaction ratings for the two candies corroborated our preference results: participants rated the Mars bar as significantly more satisfying than the lollipop, $t(242) = 12.67$, $p < .001$. Importantly, the effect size for this difference in satisfaction ratings was large ($d = .81$), suggesting that the difference is of practical as well as statistical significance. Satisfaction ratings did differ by lollipop colour, but only when participants answered the continuous satisfaction-rating questions first. Here, we found that participants rated Black lollipops significantly lower than Green lollipops ($p = .006$). Consequently, we did not give Black lollipops to participants in future experiments. Finally, we found several order effects on

preference and satisfaction ratings for the Mars bar. When people chose between the Mars bar and lollipop *before* rating their satisfaction with each, they were more likely to choose the Mars bar than participants who rated satisfaction with the individual candies first, $t(241) = 2.38$, $p = .018$. Additionally, those who answered the dichotomous preference question first rated the Mars bar higher than those who answered the continuous satisfaction questions first, $t(241) = 2.58$, $p = .011$. Perhaps having to explicitly choose between the highly-desirable Mars bar and the less-desirable lollipop inflated the Mars-lollipop difference in the subsequent questions (i.e., priming participants to think dichotomously may have led to greater differentiation between the two candies). The greater difference in satisfaction after choosing between the two candies could also be response bias to maintain consistency with the prior dichotomous choice.

Based on these findings, we chose to use Mars bars and lollipops in our main peak-end experiments. The order effects, though interesting, are unlikely to have affected our main results. In the main experiment, participants who made continuous ratings in the presence of two candies did so after seeing both candies (inviting a comparative rating). Additionally, the preference question was always asked after the satisfaction rating had been made. If anything, we would expect the Mars-lollipop difference in the main experiment to be somewhere between that of pilot participants who made the dichotomous choice first (Mars m – lollipop $m = 1.94$) and that of pilot participants who made the continuous choice first (Mars m – lollipop $m = 1.32$). It is worth noting that the mean satisfaction ratings for our candies were lower than the ratings found in the original experiment (Do et al., 2008). The mean rating for the Mars bar was 5.16, lower than the mean ceiling rating of 7 in the original experiment for the Hershey's bar. The mean satisfaction rating for the lollipop was 3.53, slightly higher than the mean rating of 3 in the original experiment) for the bubblegum in the original experiment. Despite these differences in

ratings, both candies were rated positively, and satisfaction ratings differed greatly in the predicted direction.

1. Experiment 1: Bayesian Analyses

1A. Bayesian overview and analysis specifics. To compare the relative plausibility of competing hypotheses, one must model them with a Bayesian prior. To reflect our uncertainty of the true peak-end rule effect size (given only Do et al.'s underpowered results), we chose to adopt a commonly used diffuse prior for this analysis (see Rouder, Speckman, Sun, Morey & Iverson, 2009 for a more in-depth coverage of priors and our chosen prior form, the JZS prior). We chose a one-sided prior (reflecting our prediction of a positive peak-end difference in one direction) with a width of $r = .707$. This specification of r implies that we are 50% confident that the true peak-end rule effect size (i.e., A > A_B) is somewhere between $d = 0$ and $d = .707$, and 50% confident that the true effect size is larger than $d = .707$[4]. This model of H_1 can be contrasted with a point-null model of H_0, which has all of its mass on $d = 0$.

In Bayesian analysis, the degree to which we should change our beliefs about one hypothesis relative to another is quantified by the *Bayes factor* (BF). BFs larger than 3 in favor of one hypothesis or another are generally taken to indicate meaningful evidence (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

1B. Robust Bayesian analysis. Though the use of the standard JZS prior with a medium width ($r = .707$) is both a reasonable default and standard practice, BFs can be sensitive to the priors specified. To allay potential concerns about our choice of prior, we conducted a Bayesian robustness analysis. These analyses plot BFs as a function of varying prior widths

[4] "For a symmetric Cauchy prior, the width equals the interquartile range" (Wagenmakers, 2015)

(Wagenmakerset al., 2011). Figure 1 below depicts the results of this analysis[5], and shows that for most prior widths, the evidence weighs against our peak-end rule hypotheses. It is worth noting that as prior widths get narrower, H_1 and H_0 become more similar, and obtaining evidence for H_0 becomes increasingly difficult.

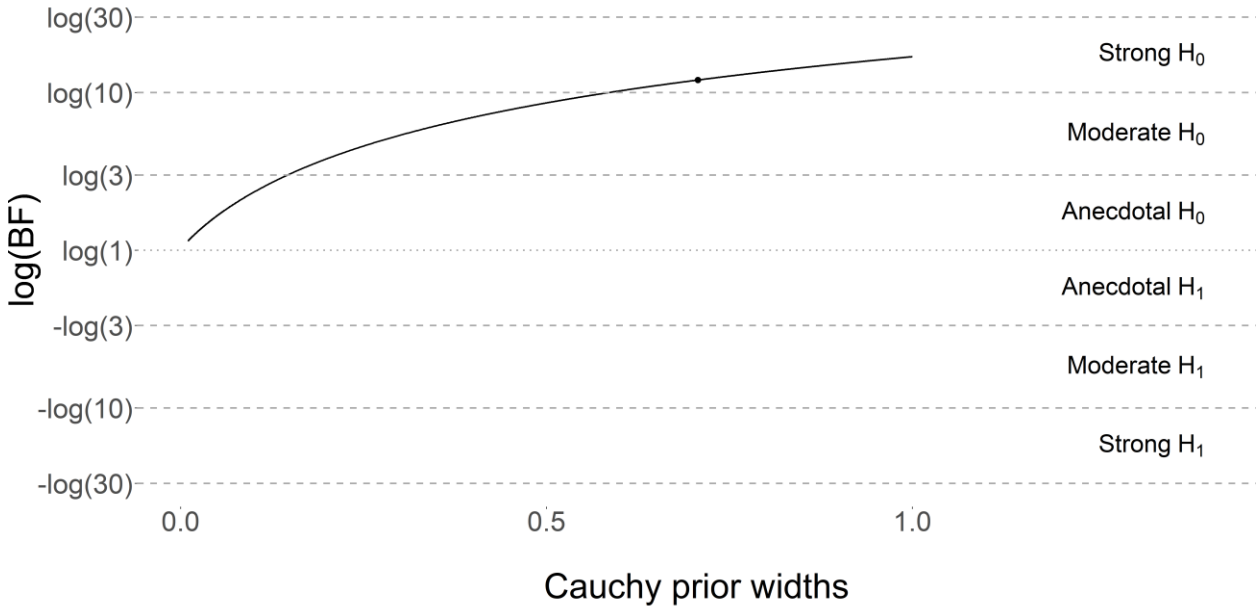

Figure 1. Bayes factor robustness plot: Experiment 1. Bayes factors plotted against various prior widths and typical standards of interpretation. H_0 = No peak-end rule effects, H_1 = Peak-end rule effect (A > A_B). The dot represents the Bayes factor at our chosen prior width of $r = .707$.

1C. Bayesian analyses using non-default priors. To examine the effects of other, non-default priors on our results, we also conducted analyses using priors that made more specific predictions about the peak-end rule[6]. For our first non-default analysis, we assumed that

[5] BF robustness plot based on R code provided by Felix Schönbrodt (2013) at https://www.nicebread.de/exploring-the-robustness-of-bayes-factors-a-convenient-plotting-function-2/ and James Grange (2015) at https://jimgrange.wordpress.com/2015/11/27/animating-robustness-check-of-bayes-factor-priors/

[6] These analyses are not standard in the literature. We refer interested readers to Dienes (2014) and Dienes & Mclatchie (2018) for more detailed discussions of informed priors, including their benefits, implications, and assumptions, and to Wiens (2017) for the R code used for these analyses.

all positive effect sizes less than or equal to the one observed by Do and colleagues (2008) were equally plausible. Thus, for this analysis the form of the prior for H_1 was a uniform distribution with a minimum of 0 (no effect) to a maximum of 1.09 (the scale difference Do and colleagues (2008) observed in their study). Using this prior we obtained a BF_{01} of 7.69 (moderate evidence for H_0). One could reasonably argue that effect sizes as large as that observed by Do and colleagues (2008) are not as plausible as smaller effect sizes. Thus, for a second non-default analysis our prior was a half-normal distribution centered on 0, with a standard deviation half the size of Do and colleagues' (2008) observed effect. With this prior specification, we observed a BF_{01} of 5 (still moderate evidence for H_0).

Though the results of our omnibus Candy combination X age group test pertain less directly to the peak-end rule, our decision to collapse across age groups relies on the assumption of no interaction. Using a Bayesian ANOVA with default specifications (see Rouder, Morey, Verhage, Swagman, & Wagenmakers, 2016 for an overview), we obtained a BF_{01} of 19.94. In other words, a model of the peak-end rule *without* a Candy combination X age interaction is 19.94 times as likely as a model with such an interaction. This analysis provides strong evidence that the peak-end rule (lack thereof) does not differ across our age groups.

2. Experiment 2: Additional Results

2A. Bayes factor robustness plot

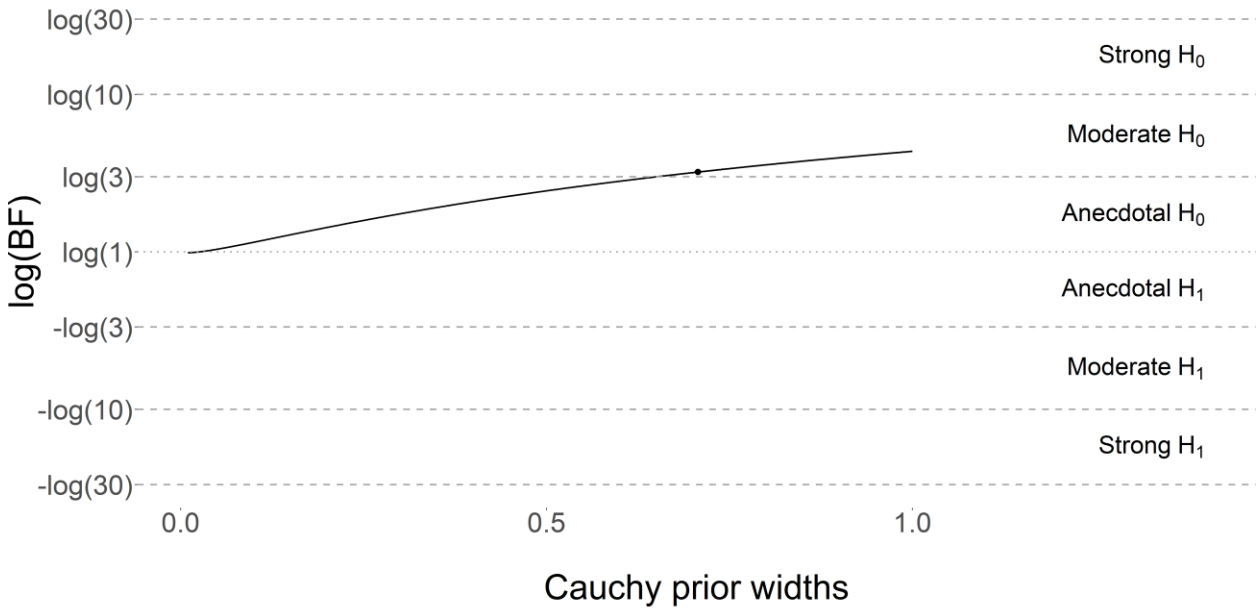

Figure 2. Bayes factor robustness plot: Experiment 2. Bayes factors plotted against various prior widths and typical standards of interpretation. H_0 = No peak-end rule effects, H_1 = Peak-end rule effect (A > A_B). The dot represents the Bayes factor at our chosen prior width of r = .707.

3. Experiment 3: Additional Results

3A. Bayes factor robustness plot

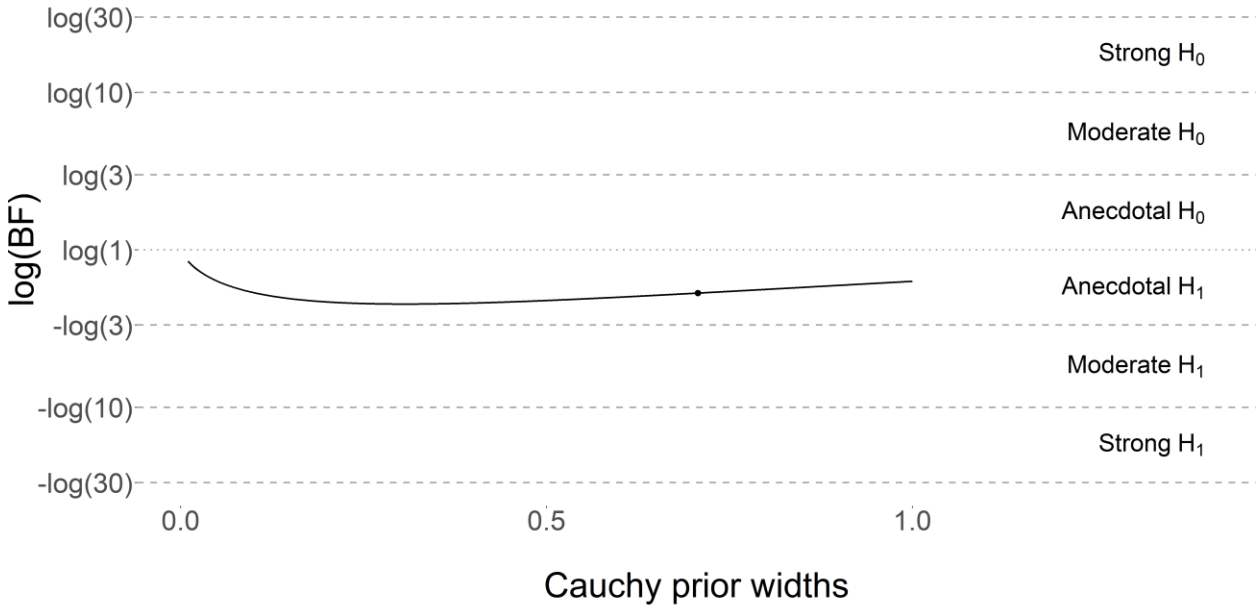

Figure 3. Bayes factor robustness plot: Experiment 3. Bayes factors plotted against various prior widths and typical standards of interpretation. H_0 = No peak-end rule effects, H_1 = Peak-end rule effect (A > A_B). The dot represents the Bayes factor at our chosen prior width of r = .707.

2B. Bayesian robust meta-analysis: Experiments 1-3

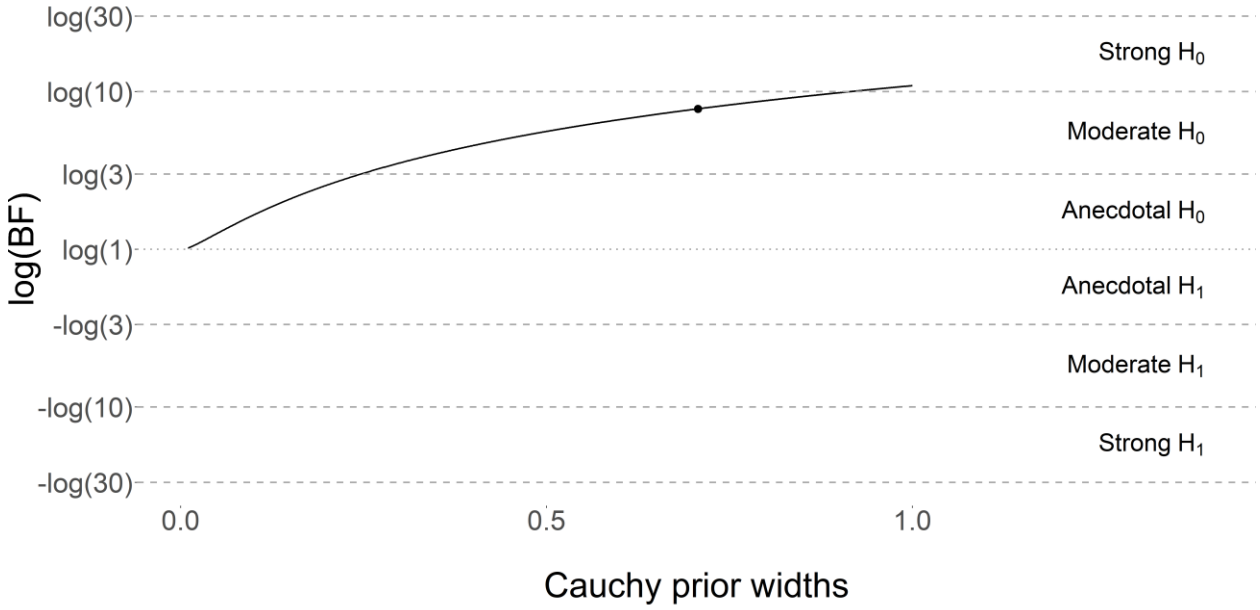

Figure 4. Bayes factor robustness plot: Experiments 1, 2 & 3 meta-analysis. Bayes factors plotted against various prior widths and typical standards of interpretation. H_0 = No peak-end rule effects, H_1 = Peak-end rule effect (A > A_B). The dot represents the Bayes factor at our chosen prior width of r = .707.