Acta Psychologica 176 (2017) 39–46

Contents lists available at ScienceDirect

Acta Psychologica

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

Sample size bias in retrospective estimates of average duration☆

Andrew R. Smith[a,*], Shanon Rule[b], Paul C. Price[c]

[a] Department of Psychology, Appalachian State University, United States
[b] Department of Psychological and Brain Sciences, University of Iowa, United States
[c] Department of Psychology, California State University, Fresno, United States

ARTICLE INFO

Keywords:
Sample size bias
Time perception
Judgments of duration
Temporal cognition

ABSTRACT

People often estimate the average duration of several events (e.g., on average, how long does it take to drive from one's home to his or her office). While there is a great deal of research investigating estimates of duration for a single event, few studies have examined estimates when people must average across numerous stimuli or events. The current studies were designed to fill this gap by examining how people's estimates of average duration were influenced by the number of stimuli being averaged (i.e., the sample size). Based on research investigating the *sample size bias*, we predicted that participants' judgments of average duration would increase as the sample size increased. Across four studies, we demonstrated a sample size bias for estimates of average duration with different judgment types (numeric estimates and comparisons), study designs (between and within-subjects), and paradigms (observing images and performing tasks). The results are consistent with the more general notion that psychological representations of magnitudes in one dimension (e.g., quantity) can influence representations of magnitudes in another dimension (e.g., duration).

1. Introduction

People often estimate the average duration of a set of similar events. For example, after grading a few papers, a college professor might estimate the average amount of time it took her to grade each paper. Or, a commuter might estimate how long, on average, it takes to drive from his home to his office. Among the reasons that such estimates of average duration are important is that they might play a critical role in planning for the future. For example, the professor might want to know how long it will take to grade the remaining papers, and the commuter might want to know if he has enough time to stop at the grocery store and still make it to work on time. However, while there has been a great deal of research investigating estimates of the duration of an individual event, very few studies have investigated estimates of the average duration of a set of events. Specifically, few studies have required that people provide estimates of average duration after experiencing numerous events of stimuli. The present research helps to close this gap in the literature.

Our focus here is on the effect of the number of events being averaged—the sample size—on estimates of average duration. Does it make a difference whether the professor is averaging the duration of two papers or ten? And if so, in what way does it make a difference? Our work is motivated by the hypothesis that people will give larger estimates of average duration as the sample size increases. We refer to this as a *sample size bias*. In the rest of this article, we present the theoretical and empirical rationale for this hypothesis followed by four experiments that empirically demonstrate the sample size bias in retrospective judgments of duration.

There are two lines of research that, taken together, suggest that there will be a sample size bias on estimates of average duration. The first consists of studies showing that estimates of duration can be biased by other nontemporal dimensions of a stimulus (e.g., Alards-Tomalin, Leboe-McGowan, Shaw, & Leboe-McGowan, 2014; Brigner, 1986; Fabbri, Cancellieri, & Natale, 2012; Matthews, Stewart, & Wearden, 2011; Xuan, Zhang, He, & Chen, 2007; for a review, see Matthews & Meck, 2016). For example, Casasanto and Boroditsky (2008) had participants estimate the amount of time that a line was displayed onscreen and found that duration estimates tended to increase as the length of the line increased. Similarly, Lu, Hodges, Zhang, and Zhang (2009; see also, Oliveri et al., 2008) had participants reproduce the duration of a numeral displayed on a computer screen for between 1000 and 5000 ms and found that as the numerical value of the numeral increased, participants' duration estimates increased as well.

☆ We would like to thank two anonymous reviewers for providing numerous references related to our studies and helpful feedback on early drafts of this manuscript. This research was supported by grants SES 12-60777 (A.R.S.) and SES 12-60642 (P.C.P.) from the National Science Foundation.

* Corresponding author at: Department of Psychology, Appalachian State University, Boone, NC 28608, United States.
E-mail address: smithar3@appstate.edu (A.R. Smith).

http://dx.doi.org/10.1016/j.actpsy.2017.03.008
Received 30 September 2016; Received in revised form 17 March 2017; Accepted 22 March 2017

The second line of research that suggests there will be a sample size bias in estimates of average duration is research on sample size bias in other domains. It has previously been shown that judgments of averages in other domains are biased by sample size. For example, in one study participants saw risk-factor information for each of several fictional employees at a company—one at a time—and then judged the risk that the average employee at the company would experience a heart attack (Price, 2001). Critically, the number of employees in the company was one, five, or nine. Price found that participants' estimates tended to increase as the number of employees in the companies increased. The sample size bias has been demonstrated across a wide variety of judgments including average judgments of the likelihood of future events (Price, 2001; Price, Smith, & Lench, 2006), estimates of the arithmetic mean of an array of numbers (Smith & Price, 2010), and judgments of the average size of groups of shapes (Price, Kimura, Smith, & Marshall, 2014). Across all of these dimensions, people's average judgments increased as the sample size increased. Previous studies investigating the sample size bias have not examined duration judgments. However, given the wide variety of domains in which the sample size bias has been observed, we expected to see a similar effect with estimates of average duration.

In addition to the theoretical support for the prediction that sample size will influence estimates of average duration, there is limited empirical support as well. In an early study investigating the influence of repetition and actual duration on people's perceptions of duration, Hintzman (1970) presented a long sequence of three-letter words to participants, with each word being presented either 1, 2, 3, or 5 times for 2, 3, 4, 5, or 6 s each time. The participants were told that each time a word was presented, it was presented for the same amount of time and their task was to estimate that single-trial exposure duration, ignoring the presentation frequency. Hintzman found that participants' duration estimates were influenced by the presentation frequency such that the more times they saw a word the longer they estimated its single-trial exposure duration to be.

In another series of related studies, Betsch, Glauer, Renkewitz, Winkler, and Sedlmeier (2010) investigated the influence of presentation frequency on estimated duration. In their studies, participants were shown a stimulus numerous times and then provided a verbal estimate of the total duration the stimulus was displayed. In general, Betsch et al. found that stimuli that were displayed more frequently were estimated to have longer total durations. For example, participants gave higher total duration estimates for items presented 8 times for 2 s each time as compared to items presented 2 times for 8 s each time.

Although the studies by Hintzman (1970) and Betsch et al. (2010) suggest that sample size will bias explicit estimates of average duration, this specific hypothesis has yet to be tested. In the studies by Hintzman (1970; see also Betsch et al., 2010, Study 3), the participants were instructed to estimate the duration of a single exposure of each item and were told that each time they saw an item it was displayed for the same amount of time. Therefore, participants could have simply thought back to the most recent exposure of the item when providing their estimate (i.e., they did not give a judgment of average duration). In order to more accurately assess the influence of sample size on estimates of average duration, participants in the current studies were not told that the items were displayed for the same amount of time during each occurrence and the participants were explicitly required to provide a judgment about the average duration of each item.

1.1. Current studies

We conducted four experiments designed as empirical demonstrations of the sample size bias in judgments of duration. In these experiments, participants estimated the average durations of items or events that varied with regard to their frequency. In all experiments, the participants were not told they would be assessing the duration of the items or events—that is, they made retrospective judgments of duration (for a discussion of retrospective and prospective judgments of duration, see Block, Hancock, & Zakay, 2010; Block & Zakay, 1997). We focused on retrospective judgment because they place a greater emphasis on memory-based cognitive processes (Block et al., 2010)—processes we speculated would be influenced by nontemporal magnitudes. Furthermore, in a number of previous studies investigating the sample size bias, participants were exposed to stimuli and made estimates about the stimuli from memory (e.g., Smith & Price, 2010, Experiment 2). We, therefore, had reason to believe that magnitude representations of sample size stored in memory could influence estimates of the average of a sample.

In Experiments 1 through 3, participants were shown a number of images that varied both in duration and sample size. In Experiment 1, after seeing the images, participants provided numeric estimates of the average duration for each image. In Experiment 2, participants indicated which of a pair of images had a longer average duration. Experiment 3 extended the findings by having participants make average duration estimates across different images. Experiment 4 further extended the findings to a new context. Rather than passively observing images, participants engaged in 24 rounds of a task and estimated the average duration of the last 2, 6, or 10 rounds. Across all of the studies, we predicted that participants' retrospective judgments of duration would increase as the sample size increased—that is, they would exhibit a sample size bias.

2. Experiment 1

Experiment 1 was an initial investigation into the influence of presentation frequency (i.e., sample size) on participants' retrospective judgments of average duration. Participants were shown a number of images that varied in terms of their average duration and sample size. After seeing the images, the participants were asked to make judgments of average duration. They were also asked to estimate the number of times they saw each image (i.e., the sample size) to test whether the participants maintained at least a rough approximation of the sample size, even though they were not explicitly instructed to do so.

2.1. Method

2.1.1. Participants

Forty-nine undergraduate students (51% women, 49% men; M_{age} = 20.20, SD_{age} = 2.29) from Appalachian State University participated as partial fulfillment of a research requirement.[1]

2.1.2. Materials and design

The stimuli presented to participants were six images from the International Picture Naming Project (IPNP; Bates et al., 2003). The specific images used were line drawings of a baby, bird, camel, crab, drum, and elephant. Each image was shown for 5 or 9 s each time it was displayed and was displayed 2, 6, or 12 times. Participants saw one image in each of the six possible combinations of duration and sample size. The particular image displayed in each combination was randomized across participants. The study was a 2 (average duration: 5 or 9) × 3 (sample size: 2, 6, or 12) within-subjects design.

2.1.3. Procedure

After reading a consent form, participants were told they would be shown a number of images—some multiple times—and would be asked questions about them at a later time. Participants were not told that they would be asked about the durations of the images or about the

[1] We had an a priori target of 50 participants for each of the first three studies. Because sign-ups were posted 1week at a time, the actual number of participants varied. In Experiment 4, because of the between-subjects design, we had a target of 180 participants.

number of times they saw each image. After reading the instructions, the images were displayed one at a time in an order that was randomized for each participant. There was a 1000 ms interval in between the display of the images.

After viewing the images, participants were again shown each image one at a time in a random order. While viewing each image, they were asked, "On average, how long was the above image displayed each time you saw it?" The participants provided their estimates in a text field with the label "seconds" following it. Participants provided their estimates for each of the six pictures in a random order. After providing all six estimates of average duration, participants were again shown each image one at a time and were asked to provide an estimate of the sample size of each image. Specifically, they were asked, "How many times was the above image displayed?" Finally, the participants were asked demographic questions (age and gender), debriefed, and excused.

2.2. Results and discussion

2.2.1. Estimates of average duration

In order to investigate the influence of actual average duration and sample size on participants' estimates, we conducted a 2 (average duration) × 3 (sample size) analysis of variance (ANOVA) on their estimates of average duration. As predicted, this analysis revealed a linear effect of sample size, $F(1, 48) = 7.90$, $p = 0.007$, $\eta_p^2 = 0.14$; participants' estimates tended to increase as the sample size increased (see Fig. 1). This analysis also revealed a main effect of average duration, $F(1, 48) = 6.53$, $p = 0.01$, $\eta_p^2 = 0.12$. Participants gave higher average duration estimates for the images displayed onscreen for 9 s relative to those displayed onscreen for 5 s, although participants tended to underestimate the duration of all images. Finally, there was no Average Duration × Sample Size interaction, $F(2, 47) = 0.73$, $p = 0.49$, $\eta_p^2 = 0.03$. The magnitude of the sample size bias was similar across the two duration conditions. In short, participants' estimates were influenced by both the actual average duration as well as the sample size.

2.2.2. Estimates of sample size

Even though the participants were not explicitly instructed to keep track of the sample size, we predicted that participants would encode this information. To test this prediction, we conducted a 2 (average duration) × 3 (sample size) ANOVA on participants' estimates of sample size. Consistent with the prediction that participants' encoded the sample size, there was a linear effect of sample size, $F(1, 48) = 177.18$, $p < 0.001$, $\eta_p^2 = 0.78$. Participants' sample size estimates increased as the sample size increased. (See Table 1 for descriptive statistics for the estimates of sample size for this and the other experiments that assessed estimates of sample size.) This analysis also revealed a main effect of average duration, $F(1, 48) = 4.37$, $p = 0.04$, $\eta_p^2 = 0.08$. Participants gave higher estimates for the items that were displayed onscreen for 9 s as compared to those displayed onscreen for 5 s. Finally, there was no interaction between average duration and sample size, $F(2, 47) = 0.23$, $p = 0.80$, $\eta_p^2 = 0.009$. This analysis suggests that participants encoded the sample size and that participants' estimates of sample size were influenced by the average duration of the images.

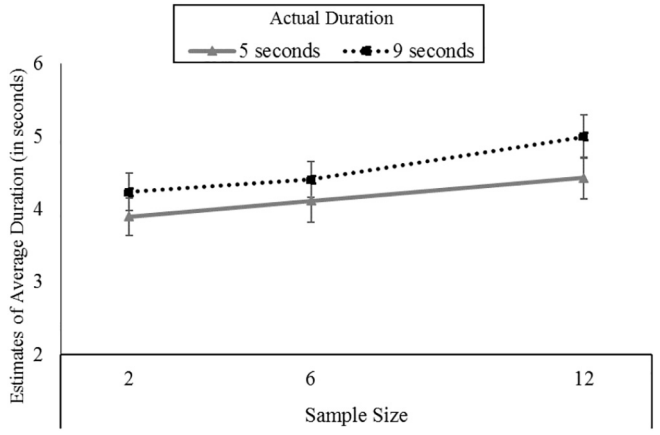

Fig. 1. Estimates of average duration as a function of actual average duration and sample size in Experiment 1. Error bars represent ± 1 *SE*.

Table 1
Means (and SDs) of sample size estimates in Experiments 1–3.

Experiment 1			
	Sample size		
Average duration	2	6	12
5 s	4.16 (2.73)	8.12 (3.09)	13.51 (6.01)
9 s	5.08 (3.51)	8.55 (2.96)	14.22 (4.98)

Experiment 2		
	Sample size	
Average duration	3	9
5 s	5.81 (3.57)	11.44 (4.10)
10 s	6.59 (4.16)	12.85 (5.17)

Experiment 3		
	Sample size	
Average duration	3	11
4 s	5.08 (3.21)	9.89 (4.58)
10 s	5.57 (5.12)	11.31 (4.54)

2.2.3. Discussion

As hypothesized, participants' estimates of average duration were influenced by the sample size. Specifically, as the sample size increased, so did participants' average duration estimates. Importantly, participants' estimates of average duration were also sensitive to the actual average durations of the images; participants gave longer duration estimates for the stimuli that were, in fact, on screen longer.

In addition to showing a sensitivity to duration information, participants' estimates of sample size were sensitive to the number of times the images were displayed. Finally, participants' estimates of sample size were also influenced by the average duration.

3. Experiment 2

Experiment 1 provided evidence that sample size can bias estimates of average duration. In Experiment 2, we sought to extend this finding by having participants make comparative judgments, rather than providing numeric estimates. This was important for two reasons. First, because there are different cognitive processes involved in making a comparison between two stimuli as opposed to making an estimate of a single stimulus, it was possible that the sample size bias would not extend to comparative judgments. Second, this study helped to test whether sample size was biasing people only at the point when they were required to produce a numeric estimate, or if sample size biased people's internal representation of average duration.

In short, Experiment 2 was not designed to distinguish between comparative judgments and numeric estimates; Experiment 2 was designed to establish the generality of the sample size bias and identify

whether the sample size biases people's internal representation of duration.

3.1. Method

3.1.1. Participants

Seventy-three undergraduate students from Appalachian State University participated as partial fulfillment of a research requirement. Although demographic information was not recorded for this experiment, the distributions of gender and age were likely quite similar to the other experiments.

3.1.2. Materials and design

Participants were presented with eight images from the IPNP (baby, bird, camel, crab, drum, elephant, fire hydrant, and ghost). Each image was shown for 5 or 10 s each time it was displayed and was displayed 3 or 9 times. Participants saw two images in each of the four possible combinations of durations and sample sizes. The particular image displayed in each combination was randomized across participants. The study was a 2 (average duration: 5 or 10) × 2 (sample size: 3 or 9) within-subjects design.

3.1.3. Procedure

Displaying the images to the participants was very similar to Experiment 1. Specifically, participants were told that they would be shown a number of pictures and would be asked questions about them at a later time, but they were not told that they would be asked about the duration the images or how many times they were displayed. The images were displayed onscreen in an order that was randomized for each participant, with a 1000 ms interval in between the display of the images.

After viewing the images, the participants were shown two images at a time—one on the right side of the screen and one on the left side. The side of the screen that the images were displayed on was randomized for each participant. While viewing each image pair, the participants were asked, “On average, which image was displayed for a longer amount of time each time you saw it?” The participants provided their answers by clicking on one of three response options: “The image on the LEFT was displayed longer each time,” “The two images were displayed for the same amount of time,” “The image on the RIGHT was displayed longer each time.” Each participant made comparative judgments about four image pairs. In two of the image pairs, both images had the same duration but different sample sizes. In the other two image pairs, both images had the same sample sizes but different durations. After making the four comparative duration judgments, the participants were shown each image one at a time and provided the same judgment of sample size as in Experiment 1. Finally, the participants were debriefed and excused.

3.2. Results and discussion

3.2.1. Estimates of duration

As mentioned earlier, the participants saw two image pairs where the average duration was held constant across the pair, but the sample size varied. By evaluating the comparative judgments between these two image pairs, we were able to assess whether the sample size influenced the participants' perceptions of duration. When comparing the two images that were displayed for 5 s, more participants indicated that the image displayed 9 times (53.4%) had a longer average duration than the image displayed 3 times (24.7%) or that the two images were displayed for the same amount of time (21.9%), $\chi^2(2) = 13.34$, $p = 0.001$. Similarly, when comparing the two images that were displayed for 10 s, more participants indicated that the image displayed 9 times (56.2%) had a longer average duration than the image displayed 3 times (21.9%) or that the two images were displayed for the same amount of time (21.9%), $\chi^2(2) = 17.12$, $p < 0.001$. In short, holding the actual average duration constant, participants were more likely to indicate that the image displayed more times was displayed onscreen longer as compared to the image displayed fewer times.

Participants also saw two image pairs where the sample size was held constant across the pair, but the duration varied. By examining these judgments, we can evaluate whether the participants were sensitive to the actual average durations of the images. When comparing the two images that were displayed 3 times, more participants indicated that the image displayed for 10 s (47.9%) was onscreen longer as compared to responding that the image displayed for 5 s (20.5%) was onscreen longer or that the two images were displayed for the same amount of time (31.6%), $\chi^2(2) = 8.33$, $p = 0.02$. Similarly, when comparing the two images that were displayed 9 times, more participants indicated that the image displayed for 10 s (53.4%) was onscreen longer as compared to responding that the image displayed for 5 s (21.9%) was onscreen longer or that the two images were displayed for the same amount of time (24.7%), $\chi^2(2) = 13.34$, $p = 0.001$. In short, holding the sample size constant, participants were more likely to correctly indicate that the image that was actually displayed onscreen longer was, in fact, displayed longer.

3.2.2. Estimates of sample size

To ensure that the participants formed at least an approximate representation of the sample size, we evaluated participants' estimates of sample size. A 2 (average duration) × 2 (sample size) ANOVA was performed on the average of the two sample size estimates for each duration-sample size combination (see Table 1). This analysis revealed a main effect of sample size, $F(1, 72) = 232.83$, $p < 0.001$, $\eta_p^2 = 0.76$. Specifically, participants gave higher estimates of sample size for the image that was displayed 9 times as compared to the image that was displayed 3 times. There was also a main effect of average duration, $F(1, 72) = 12.30$, $p = 0.001$, $\eta_p^2 = 0.15$; participants gave higher estimates of sample size for the images that were displayed 10 s each time as compared to the images that were displayed 5 s. As in Experiment 1, the average duration of the items influenced participants' estimates of sample size. Finally, there was no interaction, $F(1, 72) = 1.78$, $p = 0.19$, $\eta_p^2 = 0.02$. This analysis provides support for the notion that participants—to some extent—were aware of the approximate sample sizes of the images.

3.2.3. Discussion

The results of Experiment 2 are consistent with Experiment 1; participants' perceptions of average duration were influenced by the sample size of the stimuli. Importantly, this occurred when participants were making comparisons between images. Because participants never provided numeric estimates of duration, this rules out the explanation that sample size only influences the numeric estimates people give, but not their internal representation of average duration. As in Experiment 1, the participants were also sensitive to the actual average durations of the images.

4. Experiment 3

The previous experiments demonstrate that sample size influences estimates of average duration. However, in the previous experiments, participants were averaging across numerous instances of the same stimulus. Because the duration that an image was displayed each time was constant, participants could in principle recall one instance and provide a judgment regarding that single instance despite being instructed to average across the instances. In Experiment 3, we addressed this issue by requiring participants to average across different items.

Experiment 3 also served to test the generality of the sample size bias. Previous research has made a distinction between estimates of a specific item and estimates of a category of items, and found that these two types of estimates are influenced by different factors. For example,

Manis, Shedler, Jonides, and Nelson (1993) found that people's estimates of the frequency of a category of items (e.g., the number of female names in a list) were influenced by the availability of instances from the category. However, estimates of the frequency of specific instances (e.g., the number of times "Mary" was repeated in the list) were not influenced by availability. Given these differences, we tested whether the sample size bias would be present when participants made estimates across a category of items (in addition to when making estimates of a specific item as demonstrated in the previous experiments).

4.1. Method

4.1.1. Participants

Sixty-one participants (66.7% women, 32.8% men, 1.6% did not respond; M_{age} = 19.05, SD_{age} = 1.48) from Appalachian State University participated as partial fulfillment of a research requirement.

4.1.2. Materials and design

In this experiment, participants were shown eight slightly modified images from the IPNP. These eight images created four images pairs by combining images from the same general category (pair 1 = boy and girl; pair 2 = pot and tea pot; pair 3 = camel and elephant; pair 4 = cactus and wheat). Furthermore, the background of each pair was painted a particular color (pair 1 = blue; pair 2 = green; pair 3 = red; pair 4 = yellow). Each image pair was presented to the participants either 3 (2 for one image, 1 for the other) or 11 (6 for one image, 5 for the other) times. The pairs were displayed for 4 or 10 s each time they were onscreen. Participants saw one pair in each of the four possible combinations of durations and sample sizes. The particular image pair in each combination was randomized across participants. The study was a 2 (average duration: 4 or 10) × 2 (sample size: 3 or 11) within-subjects design.

4.1.3. Procedure

As in the previous studies, the participants were told they would be shown a number of pictures and would be asked questions about them at a later time, but they were not told they would be asked about the duration the images were displayed. The images were displayed onscreen in an order that was randomized for each participant with a 1000 ms interval in between the display of the images.

After viewing the images, the participants were shown a colored square (blue, green, red, or yellow) and were asked, "On average, how long were the [blue, green, red, yellow] images displayed on screen each time you saw them?" The participants provided their estimates in a text field with the label "seconds" following it. The questions about the four image colors were presented in a random order. After making the estimates of duration, the participants were again shown each colored square and asked to estimate the sample size. Specifically, they were asked, "How many times were the [blue, green, red, yellow] images displayed on screen?" Finally, the participants were asked demographic questions (age and gender), debriefed, and excused.

4.2. Results and discussion

4.2.1. Estimates of average duration

In order to assess the influence of the sample size on participants' duration estimates, a 2 (average duration) × 2 (sample size) ANOVA was conducted. There was a main effect of sample size, $F(1, 60) = 10.81$, $p = 0.002$, $\eta_p^2 = 0.15$. Consistent with our prediction, participants' estimates increased as the sample size increased (see Fig. 2). This analysis also revealed a main effect of average duration, $F(1, 60) = 19.72$, $p < 0.001$, $\eta_p^2 = 0.25$. Participants gave higher average duration estimates for the images displayed onscreen for 10 s as compared to those displayed onscreen for 4 s. Finally, there was a significant Average Duration × Sample Size interaction, $F(1, 60) = 5.64$, $p = 0.02$, $\eta_p^2 = 0.09$. Follow-up contrasts revealed that when the duration was 4 s, participants gave higher average duration estimates when the sample size was 11 relative to when the sample size was 2, $F(1, 60) = 4.43$, $p = 0.04$, $\eta_p^2 = 0.07$. Similarly, when the duration was 10 s, participants gave higher average duration estimates when the sample size was 11 relative to when the sample size was 3, $F(1, 60) = 10.48$, $p = 0.002$, $\eta_p^2 = 0.15$.

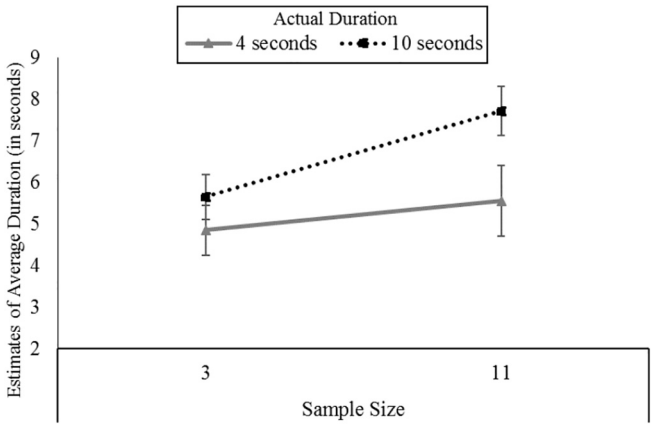

Fig. 2. Participants' estimates of average duration as a function of actual average duration and sample size Experiment 3. Error bars represent ± 1 *SE*.

4.2.2. Estimates of sample size

As in the previous experiments, we evaluated participants' judgment of sample size to ensure the participants formed an approximate representation of this dimension. A 2 (average duration) × 2 (sample size) ANOVA on participants' sample size estimates revealed a main effect of sample size, $F(1, 60) = 93.46$, $p < 0.001$, $\eta_p^2 = 0.61$ (see Table 1). Participants gave higher sample size estimates for the pairs that were displayed 11 times as compared to 3 times. There was a marginally significant main effect of average duration, $F(1, 60) = 3.81$, $p = 0.06$, $\eta_p^2 = 0.06$. Participants' sample size estimates were somewhat higher for the pairs that were displayed for 10 s as compared to the pairs that were displayed for 4 s. Finally, there was not an Average Duration × Sample Size interaction, $F(1, 60) = 1.59$, $p = 0.21$, $\eta_p^2 = 0.03$.

4.2.3. Discussion

As in the previous experiments, participants' estimates of average duration were influenced by the number of items they averaged across. Specifically, as the sample size increased, so did participants' judgment of average duration. Importantly, this occurred even when participants were averaging across different items (e.g., an image of a camel and an image of an elephant). Participants' duration estimates were also influenced by the actual durations of the images. That is, they gave higher average duration estimates for the images that were, in fact, onscreen for more time.

5. Experiment 4

Experiments 1 through 3 lend support to the notion that sample size influences people's estimates of average duration. However, the previous studies relied on the same paradigm. That is, participants saw numerous images and estimated the average duration that each image was displayed. In Experiment 4, we tested for a sample size bias using a different paradigm. Specifically, participants performed 24 rounds of a task and then were asked to think back to the last 2, 6, or 10 rounds and estimate the average duration of the rounds. In addition to generalizing to a new situation, this study also differed from the previous studies in a few important ways. First, the duration of each round was determined by the participant (i.e., it was the amount of time it took them to

complete the task) and not controlled by the experimenter. Second, the experienced sample size (i.e., the number of rounds of the task) was the same for all the participants, but they were asked to average across different sample sizes.

5.1. Method

5.1.1. Participants

One hundred and seventy-two undergraduate students (73.3% women, 26.7% men; M_{age} = 19.15, SD_{age} = 2.13) from Appalachian State University participated as partial fulfillment of a research requirement.

5.1.2. Materials and design

Participants in this study completed numerous rounds of the Columbia Card Task (CCT; Figner, Mackinlay, Wilkening, & Weber, 2009). The CCT is a computer-based behavioral risk-taking task where participants try to earn as many points as possible by turning over virtual cards (for a complete description of the task, see Figner et al., 2009). One important feature of the CCT is that there are two different versions—a hot and cold version. In the hot version, participants are presented with an array of 32 virtual cards and, one at a time, pick the card they would like to turn over. When they finish selecting cards, they move to the next round. In the cold version, participants are also presented with an array of 32 virtual cards. However, participants simply select a number from 0 to 32 that indicates the number of cards they would like to turn over. Once they make their selection, they move to the next rounds. Because the cold version requires only one choice per round, the average amount of time it takes participants to complete a round is much shorter in the cold version as compared to the hot version.[2]

This study was a 2 (CCT version: hot or cold) × 3 (sample size: 2, 6, or 10) between-subjects design. Note that the between-subjects design contrasts with the within-subjects design of Experiments 1 through 3 and most previous research on the sample size bias (Price, 2001; Price et al., 2006; Price et al., 2014; Smith & Price, 2010).

5.1.3. Procedure

As part of an unrelated study,[3] participants were randomly assigned to complete 24 rounds of either the hot or the cold version of the CCT. After completing the 24 rounds, the participants made an average duration judgment for the last 2, 6, or 10 rounds. Specifically, participants were asked to, “Think back to the last [2, 6, 10] rounds of the card game. On average, how long do you think it took you to go through each round?” After providing their estimate, the participants completed the remaining portion of the unrelated study. Finally, the participants were asked demographic questions (age and gender), debriefed, and excused.

5.2. Results and discussion

5.2.1. Estimates of average duration

Responses from eight participants were dropped—three because the participants either misunderstood their task or their responses were typos, and five because they were > 3 SDs above the mean.[4] To investigate the influence of sample size on participants' estimates of average duration, we conducted a 2 (CCT version: hot or cold) × 3 (sample size: 2, 6, or 10) ANOVA. As predicted, there was a main effect of sample size, $F(2, 158) = 3.49$, $p = 0.03$, $\eta_p^2 = 0.04$. Participants' duration estimates were sensitive to the size of the sample they averaged across (see Fig. 3). This analysis also revealed a main effect of CCT version, $F(1, 158) = 51.11$, $p < 0.001$, $\eta_p^2 = 0.24$. Participants gave higher average duration estimates when they went through the hot version as compared to the cold version. Finally, there was no interaction between average duration and sample size, $F(2, 158) = 1.27$, $p = 0.28$, $\eta_p^2 = 0.02$.

While the above analysis tested for a difference between estimates across the three sample sizes, it did not specifically test for a linear increase in average duration estimates as a function of sample size. Therefore, we conducted a regression analysis using the CCT version, the sample size participants were asked to consider, and the interaction term to predict participants' average duration estimate. As predicted—and consistent with the above analysis—the sample size significantly predicted participants average duration estimates, $b = 0.46$, $t = 2.61$, $p = 0.01$. As the sample size increased, so did participants' estimates of average duration. CCT version also predicted participants' average duration estimates, $b = -8.20$, $t = 7.20$, $p < 0.001$. Participants who went through the hot version gave higher estimates than participants who went through the cold version. Finally, the interaction term was not a significant predictor, $b = -0.55$, $t = 1.56$, $p = 0.12$.

Taken together, both of the above analyses support the conclusion that people's estimates of average duration are influenced by the sample size.[5] As in the previous experiments, participants' duration estimates were also influenced by the actual durations. Specifically, participants who went through the hot version of the CCT provided longer average duration estimates than participants who went through the cold version of the CCT.

6. General discussion

The current studies were designed to test the prediction that the size of a sample would influence people's estimates of average duration in a retrospective judgment paradigm. The results of all four experiments support this prediction. Across these experiments, the sample size bias was found when participants made numeric estimates of average duration (Experiments 1, 3, and 4) as well as when they made comparisons of different stimuli (Experiment 2). We observed the effect when participants made duration estimates about passively-experienced events (Experiments 1–3) and when they estimated the average duration of tasks they engaged in (Experiment 4). Furthermore, there was a sample size bias when using both within-subjects (Experiments 1–3) and between-subjects (Experiment 4) designs. And finally, we observed the effect when the sample size was explicitly mentioned (Experiment 4) and when it was not mentioned (Experiments 1–3). The

[2] The Columbia Card Task (CCT) program did not record how long participants took to complete each round. Therefore, we do not know the actual duration of the rounds. However, the program did record the total time it took for participants to complete their entire task (i.e., read the instructions, complete 3 practice rounds, and complete the 24 test rounds). Consistent with the notion that the average duration of each round was longer when in the hot vs. cold version of the CCT, the total duration to complete the task was significantly longer for the participants who went through the hot version (M = 9.89min, SD = 2.83) as compared to the cold version (M = 6.04min, SD = 1.63), $t(139) = 9.87$, $p < 0.001$, $d = 1.73$.

[3] The unrelated study investigated the relationship between risk-taking behavior as measured by the Columbia Card Task and numerous individual difference measures (e.g., anxiety sensitivity, optimism, numeracy).

[4] Inclusion of the five responses that were > 3 SDs above the mean did not substantially change the results of the analyses. Specifically, a 2 (CCT version: hot or cold) × 3 (sample size: 2, 6, or 10) ANOVA found the predicted a main effect of sample size, $F(2, 163) = 3.77$, $p = 0.03$, $\eta_p^2 = 0.04$. This analysis also revealed a main effect of CCT version, $F(1, 163) = 35.31$, $p < 0.001$, $\eta_p^2 = 0.18$. Participants gave higher average duration estimates when they went through the hot version as compared to the cold version. Finally, there was no Average Duration × Sample Size interaction, $F(2, 163) = 1.70$, $p = 0.19$, $\eta_p^2 = 0.02$.

[5] It is possible that the differences in duration judgments across the three sample size conditions could reflect differences in the actual time the participants took to complete the rounds. While possible, this is unlikely because participants were randomly assigned to the sample size conditions after they completed the card task. Also, the three sample size conditions did not differ in the total amount of time taken to complete the CCT, $F(2, 165) = 0.26$, $p = 0.79$, the points earned during the task, $F(2, 165) = 0.13$, $p = 0.88$, or the number of cards they turned over during the task, $F(2, 165) = 1.10$, $p = 0.34$.

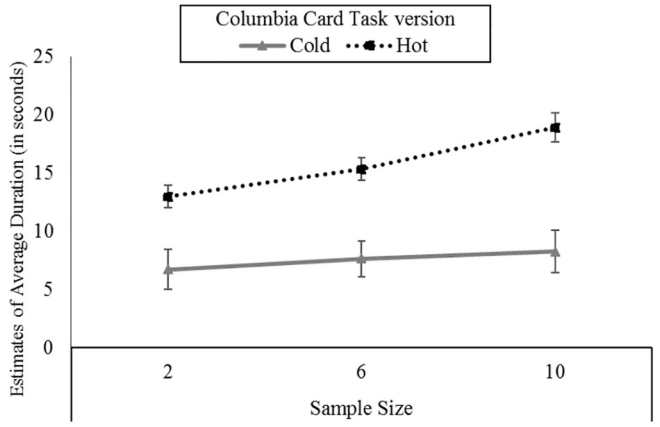

Fig. 3. Estimates of average duration as a function of sample size and Columbia Card Task version (Cold vs. Hot) in Experiment 4. Error bars represent ± 1 *SE*.

results of the current studies are consistent with previous research on the sample size bias (e.g., Price et al., 2006; Price et al., 2014; Smith & Price, 2010) and research demonstrating that nontemporal magnitudes can influence duration estimates (e.g., Brigner, 1986; Matthews et al., 2011; Oliveri et al., 2008; Xuan et al., 2007). It is important to note that the previous research investigating the influence of nontemporal magnitudes on judgments of duration generally required participants to make prospective judgments while we had participants make retrospective judgment. Therefore, it is unclear whether the same pattern of results would occur using a prospective judgment paradigm. Furthermore, the current studies were simply designed to empirically demonstrate the sample size bias in different contexts, but did not test specific mechanisms producing the effect. While the current studies were not designed to identify the precise mechanism, in the following section we speculate as to how sample size influences retrospective estimates of average duration.

When explaining how presentation frequency can bias estimates of total duration, Betsch et al. (2010) suggested that as presentation frequency increases, so does the strength of the memory trace (see also, Dougherty, Gettys, & Ogden, 1999). Because people may use the strength of the memory trace as a cue when estimating total duration, increasing presentation frequency increases perceived total duration. When applied to the current studies, it is possible that the increase in memory trace strength also influences perceptions of average duration. While the memory trace account can explain the results of some of our studies, it has trouble with Experiment 4. Because all participants completed the task exactly 24 times, the memory trace strength of these experiences should be, on average, similar across the sample size conditions—suggesting that something other than memory trace is the mediating variable.

In a recent review, Matthews and Meck (2016) described numerous accounts of why the magnitude of a stimulus influences duration estimates in prospective time judgments. The first account focuses on the pacemaker-accumulator framework (Gibbon, Church, & Meck, 1984; Meck & Church, 1983; Treisman, 1963). According to this framework, to form a representation of duration, pulses from a pacemaker are collected by an accumulator. The collected pulses represent the experienced duration and this representation is compared to a standard stored in long-term memory. Several researchers have argued that nontemporal dimensions of a stimulus can increase the rate of the pacemaker (e.g., Matthews, 2011; Penney, Gibbon, & Meck, 2000), thereby increasing the perceived duration when making prospective judgments. For example, a number of studies have found that rapid repetitions (e.g., visual flicker) can speed up the pacemaker producing longer experiences of subjective duration (e.g., Droit-Volet & Wearden, 2002; Ortega & López, 2008). While those studies investigated rapid repetitions and used temporal bisection task, perhaps in the current studies as a stimulus was observed more frequently (i.e., as the sample size increased), this increased the rate of the pacemaker. If the speed of the pacemaker increases, subjective duration estimates will likely also increase.

Another account explaining how nontemporal properties can influence perceived duration has to do with the efficiency of neural coding and suggests that "…the amount of neural energy required to represent a stimulus is proportional to, or at least influences, the subject duration…" (Eagleman & Pariyadath, 2009, p. 1842). According to this account, anything that results in a larger neural response (e.g., an increase in size or brightness) will result in a longer perceived duration. If an increase in presentation frequency increases the neural response, perceptions of average duration may be increased. While this is possible, repeated presentations of a stimulus have been found to lead to a decrease in perceptions of a single-event duration (e.g., Matthews, 2011; Rose & Summers, 1995) presumably because repeated presentations are more easily processed than novel presentations.

A final account suggests that many nontemporal properties (e.g., space, numerical magnitude, numerosity) affect perceptions of duration because there is overlap in the neural representation these magnitudes (Walsh, 2003; see also, Casasanto & Boroditsky, 2008; Conson, Cinque, Barbarulo, & Trojano, 2008; Droit-Volet & Coull, 2015; Oliveri et al., 2008; Pinel, Piazza, Le Bihan, & Dehaene, 2004; Xuan et al., 2007). Because of this common system, measurements of one dimension (e.g., duration) can be influenced by, or partly based on magnitudes of different dimensions (e.g., size, numerosity). According to this account, representations of frequency and duration are processed by a similar neural system. Therefore, perceptions of average duration are somewhat based on perceptions of frequency. This can happen both when presentation frequency is explicitly mentioned (as in our Experiment 4) and when it is not (as in the other experiments) because people automatically encode the approximate frequency of stimuli (Hasher & Zacks, 1979; Naveh-Benjamin & Jonides, 1986). The results of the current studies seem best explained by the common magnitude system account, but it is important to note that these studies were designed to demonstrate the sample size bias in estimates of duration and test for potential moderators (e.g., type of task, type of judgment, type of research design) and not designed to distinguish between these different theoretical accounts.

Another reason that we favor the common magnitude system account is that, in addition to being able to explain the results of the current studies, it provides an explanation for the previous research on the sample size bias that is outside the domain of duration estimation. As described earlier, the sample size bias has been observed when participants estimated the average heart attack risk of employees at hypothetical companies (Price, 2001), the arithmetic mean of a set of numbers (Smith & Price, 2010), and the average size of groups of shapes (Price et al., 2014). Given the generality of the effect, we favor an explanation that is similarly general, rather than an explanation that is specific to duration judgments. That being said, it is possible that the sample size bias is driven by different mechanisms across these different domains. Therefore, future research is needed to establish the exact mechanism in each decision context.

Finally, it is worth noting that in the current studies, participants were explicitly required to provide estimates of average duration. There are many estimates that do not explicitly request that people provide an average, but are in fact, estimates of average duration. For example, asking a friend how long it takes for her to get from her house to work could be thought of asking how long, on average, the trip generally takes. It is possible that judgments that do not explicitly require averaging also exhibit a sample size bias. If the worker has made the trip from her house to work many times and can easily recall numerous instances of the drive, she might estimate the duration to be longer than a worker who has recently joined the company and has only made the drive a few times. While this is possible, further research is needed to identify if estimates like those described are, in fact, influenced by

sample size.

6.1. Conclusion

The sample size bias has been observed across a growing lists of domains, including risk judgments, estimates of the arithmetic mean, and estimates of size. The four studies described in this manuscript add to this list by demonstrating a sample size bias when participants made retrospective estimates of average duration—a bias that was observed across a variety of tasks, judgment types, and research designs. In addition to demonstrating the sample size bias in a new context, the current research also provides directions for future research, such as identifying the mechanism producing the sample size bias and how this bias may affect people in real-world situations.

References

Alards-Tomalin, D., Leboe-McGowan, J. P., Shaw, J. M., & Leboe-McGowan, L. C. (2014). The effects of numerical magnitude, size, and color saturation on perceived interval duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 555–566. http://dx.doi.org/10.1037/a0035031.

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., ... Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, *10*(2), 344–380. http://dx.doi.org/10.3758/BF03196494.

Betsch, T., Glauer, M., Renkewitz, F., Winkler, I., & Sedlmeier, P. (2010). Encoding, storage and judgment of experienced frequency and duration. *Judgment and Decision Making*, *5*(5), 347–364.

Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica*, *134*(3), 330–343. http://dx.doi.org/10.1016/j.actpsy.2010.03.006.

Block, R. A., & Zakay, D. (1997). Prospective and retrospective duration judgments: A meta-analytic review. *Psychonomic Bulletin & Review*, *4*(2), 184–197. http://dx.doi.org/10.3758/BF03209393.

Brigner, W. L. (1986). Effect of perceived brightness on perceived time. *Perceptual and Motor Skills*, *63*, 427–430. http://dx.doi.org/10.2466/pms.1986.63.2.427.

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, *106*(2), 579–593. http://dx.doi.org/10.1016/j.cognition.2007.03.004.

Conson, M., Cinque, F., Barbarulo, A. M., & Trojano, L. (2008). A common processing system for duration, order and spatial information: Evidence from a time estimation task. *Experimental Brain Research*, *187*(2), 267–274. http://dx.doi.org/10.1007/s00221-008-1300-5.

Dougherty, M. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*(1), 180–209. http://dx.doi.org/10.1037/0033-295X.106.1.180.

Droit-Volet, S., & Coull, J. (2015). The developmental emergence of the mental time-line: Spatial and numerical distortion of time judgement. *PloS One*, *10*(7), e0130465.

Droit-Volet, S., & Wearden, J. (2002). Speeding up an internal clock in children? Effects of visual flicker on subjective duration. *Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *55*(3), 193–211. http://dx.doi.org/10.1080/02724990143000252.

Eagleman, D. M., & Pariyadath, V. (2009). Is subjective duration a signature of coding efficiency? *Philosophical Transactions of the Royal Society of London B*, *364*(1525), 1841–1851. http://dx.doi.org/10.1098/rstb.2009.0026.

Fabbri, M., Cancellieri, J., & Natale, V. (2012). The A Theory of Magnitude (ATOM) model in temporal perception and reproduction tasks. *Acta Psychologica*, *139*(1), 111–123. http://dx.doi.org/10.1016/j.actpsy.2011.09.006.

Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and deliberative processes in risky choice: Age differences in risk taking in the Columbia Card Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 709–730. http://dx.doi.org/10.1037/a0014983.

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. *Annals of the New York Academy of Sciences*, *423*, 52–77. http://dx.doi.org/10.1111/j.1749-6632.1984.tb23417.x.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *108*(3), 356–388. http://dx.doi.org/10.1037/0096-3445.108.3.356.

Hintzman, D. L. (1970). Effects of repetition and exposure duration on memory. *Journal of Experimental Psychology*, *83*(3), 435–444. http://dx.doi.org/10.1037/h0028865.

Lu, A., Hodges, B., Zhang, J., & Zhang, J. X. (2009). Contextual effects on number–time interaction. *Cognition*, *113*(1), 117–122. http://dx.doi.org/10.1016/j.cognition.2009.07.001.

Manis, M., Shedler, J., Jonides, J., & Nelson, T. E. (1993). Availability heuristic in judgments of set size and frequency of occurrence. *Journal of Personality and Social Psychology*, *65*(3), 448–457. http://dx.doi.org/10.1037/0022-3514.65.3.448.

Matthews, W. J. (2011). Stimulus repetition and the perception of time: The effects of prior exposure on temporal discrimination, judgment, and production. *PLoS ONE*, *6*e19815. http://dx.doi.org/10.1371/journal.pone.0019815.

Matthews, W. J., & Meck, W. H. (2016). Temporal cognition: Connecting subjective time to perception, attention, and memory. *Psychological Bulletin*. http://dx.doi.org/10.1037/bul0000045 (Advance online publication).

Matthews, W. J., Stewart, N., & Wearden, J. H. (2011). Stimulus intensity and the perception of duration. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 303–313. http://dx.doi.org/10.1037/a0019961.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(3), 320–334. http://dx.doi.org/10.1037/0097-7403.9.3.320.

Naveh-Benjamin, M., & Jonides, J. (1986). On the automaticity of frequency coding: Effects of competing task load, encoding strategy, and intention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 378–386. http://dx.doi.org/10.1037/0278-7393.12.3.378.

Oliveri, M., Vicario, C. M., Salerno, S., Koch, G., Turriziani, P., Mangano, R., ... Caltagirone, C. (2008). Perceiving numbers alters time perception. *Neuroscience Letters*, *438*, 308–311. http://dx.doi.org/10.1016/j.neulet.2008.04.051.

Ortega, L., & López, F. (2008). Effects of visual flicker on subjective time in a temporal bisection task. *Behavioural Processes*, *78*(3), 380–386. http://dx.doi.org/10.1016/j.beproc.2008.02.004.

Penney, T. B., Gibbon, J., & Meck, W. H. (2000). Differential effects of auditory and visual signals on clock speed and temporal memory. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(6), 1770–1787. http://dx.doi.org/10.1037/0096-1523.26.6.1770.

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgment. *Neuron*, *41*, 983–993. http://dx.doi.org/10.1016/S0896-6273(04)00107-2.

Price, P. C. (2001). A group size effect on personal risk judgments: Implications for unrealistic optimism. *Memory & Cognition*, *29*(4), 578–586. http://dx.doi.org/10.3758/BF03200459.

Price, P. C., Kimura, N. M., Smith, A. R., & Marshall, L. D. (2014). Sample size bias in judgments of perceptual averages. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1321–1331. http://dx.doi.org/10.1037/a0036576.

Price, P. C., Smith, A. R., & Lench, H. C. (2006). The effect of target group size on risk judgments and comparative optimism: The more, the riskier. *Journal of Personality and Social Psychology*, *90*(3), 382–398. http://dx.doi.org/10.1037/0022-3514.90.3.382.

Rose, D., & Summers, J. (1995). Duration illusions in a train of visual stimuli. *Perception*, *24*(10), 1177–1187. http://dx.doi.org/10.1068/p241177.

Smith, A. R., & Price, P. C. (2010). Sample size bias in the estimation of means. *Psychonomic Bulletin & Review*, *17*(4), 499–503. http://dx.doi.org/10.3758/PBR.17.4.499.

Treisman, M. (1963). Temporal discrimination and the indifference interval: Implications for a model of the 'internal clock'. *Psychological Monographs*, *77*(13), 1–31. http://dx.doi.org/10.1037/h0093864.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488. http://dx.doi.org/10.1016/j.tics.2003.09.002.

Xuan, B., Zhang, D., He, S., & Chen, X. (2007). Larger stimuli are judged to last longer. *Journal of Vision*, *7*(10), 1–5. http://dx.doi.org/10.1167/7.10.2.