Schweickert, R., Han, H. J., Yamaguchi, M., & Fortin, C. (2014). Estimating averages from distributions of tone durations. *Attention, Perception, & Psychophysics*, *76*, 605-620. doi: 10.3758/s13414-013-0591-1

Estimating Averages from Distributions of Tone Durations

Richard Schweickert[1]

Hye Joo Han[1]

Motonori Yamaguchi[2]

Claudette Fortin[3]

[1] Purdue University, West Lafayette, IN, USA

[2] Edge Hill University, Ormskirk, UK

[3] Université Laval, Québec, Canada

Address for correspondence:

Richard Schweickert

Department of Psychological Sciences

Purdue University, West Lafayette, IN, USA

Email: swike@psych.purdue.edu

Phone: 765-494-7986

Abstract

We examined whether estimating average duration was influenced by distribution peak location. We presented participants with samples of various tone durations. We then presented comparison tone durations. Participants judged whether each comparison duration was longer than the average sample duration. Estimates of averages were inferred from psychophysical functions. Durations were sampled from three distributions, one positively skewed, one symmetric, and one negatively skewed. In Experiment 1, every participant was presented with every distribution. Estimates of averages were unbiased for the symmetric distribution, but biased toward the long tail of each skewed distribution. This would occur if participants combined the sample to be judged with previous irrelevant samples or with comparison durations. In Experiment 2 each participant was presented with samples from only one distribution. Estimates of averages were still biased toward long tails of skewed distributions. This would occur if participants combined the sample to be judged with comparison durations, which were the same for the three distributions. In Experiment 3, each participant was presented with only one distribution and each distribution was tested with its own comparison durations, selected as percentiles of the distribution. Estimates were accurate for the smallest population mean (positively skewed distribution), but underestimated larger means. Results are explained by subjective shortening of durations in memory, with a simple equation of Scalar Timing Theory. It correctly predicts two results: Estimated averages are a linear function of stimulus means, variances are a linear function of squared stimulus means. Neither prediction is dependent on skewness of stimulus durations.

Estimating Averages from Distributions of Tone Durations

People make scheduling decisions throughout a day, from a minor choice to insert a flash drive while waiting for a page to print to a serious choice of when to apply brakes. Good scheduling requires good estimates of event durations. Presumably for events one has experienced, some estimate of central tendency of the duration can be extracted (Jarvstad, Rushton, Warren & Hahn, 2012). People are fairly good at estimating the central tendency of various aspects of stimuli, for example, size (Chong & Treisman, 2003), orientation (Parkes, Lund, Angelucci, Solomon & Morgan, 2001), or even high-level features such as emotions in faces (Haberman & Whitney, 2009). Alvarez (2011) and Bauer (2009a) provide brief reviews. Little is known about averaging of tones or their durations.

Estimating the Average of a Set of Durations

In what Albrecht, Scholl and Chun (2012) say is the first study of central tendency estimation with tones, participants estimated average pitch, and in the study of Piazza, Sweeny, Wessel, Silver and Whitney (2013), participants estimated average tone frequency. In one of the few experiments on central tendency estimation of durations, Baker (1962) showed participants a series of visual signals separated by various temporal intervals. At the end of the series, participants generated two more signals at the times they expected further signals would occur. Intervals to the generated signals were close to the average of presented temporal intervals. This averaging took place without explicit instructions.

Although averaging *per se* has been little studied for durations, relevant results come from bisection experiments. For bisection of some stimulus dimensions, participants were explicitly instructed to average two reference stimuli. But for duration bisection, instructions have been to judge which of two reference durations is closer to a comparison duration, for a series of comparison durations. The subjective midpoint of the two reference durations is then calculated

from the judgment data. Brown, McCormack, Smith, and Stewart (2005) showed that the distribution of the comparison durations had an influence on the location of the subjective midpoint. Two tones were presented four times at the beginning of the experimental session. These two tones represented the shortest and longest tones of a set of tones. Then, in each experimental trial, a tone of intermediate duration was presented and the participant had to decide whether it was more similar to the short or to the long tone. Distribution of duration values in the set of tones influenced bisection performance, with more positively skewed distributions producing lower bisection points. The effect of the distribution was greater when the ratio of the largest to smallest duration was greater. Results could be accounted for by a Range theory, inspired by the Range Frequency Theory of Parducci and colleagues (Parducci, 1965; Parducci & Perrett, 1971), which states that the judgment of a particular item depends on the range of the items and the position of the item within the range. The theory accounted for the influence of stimulus distribution in a temporal bisection task, as well as in a tone frequency bisection task. Although bisection in this form is not the same as averaging, results suggest that averaging would be influenced by the distribution of durations and their range.

Bisection experiments are informative about two reference durations, but sets larger than two are experienced in everyday life, put a larger demand on memory, and may lead to different results. Thus, in the present study, we presented tones of various durations to obtain participants' estimates of the average duration. After a block of stimulus durations was presented, comparison tones of different durations were presented. Participants judged whether each comparison duration was longer than the average of the block of stimulus durations. Two sets of durations were employed, the stimulus durations and the comparison durations. The stimulus durations to be averaged form the more basic set, so in the initial experiments these were manipulated while

comparison durations were not. Stimulus durations were drawn from one of three distributions, symmetric, positively skewed or negatively skewed.

Influences of Distributional Properties on Estimate of Average

Durations of a daily life event are typically positively skewed, bounded below by 0 and occasionally very long. Estimates of central tendency of a skewed distribution could easily be biased. Findings for skewed distributions differ, depending on type of material, presentation and response. In an early study by Spencer (1961) stimuli in one condition were sets of 10 or 20 two digit numbers, and in another were 10 or 20 points on graph paper. Spencer found that accuracy in estimating central tendency was "remarkably high," better when the standard deviations of the stimuli were small, and better with the graphical material. In a follow up, Spencer (1963) presented both normal and skewed distributions, numerically and graphically. For both materials, distances between estimates and means were greater when distributions were skewed. With skewed distributions of numbers, but not with graphed points, estimates were biased toward the long tail.

In another experiment on skewed distributions of numbers, Malmi and Samson (1983) presented participants with three digit integers said to be SAT scores of individuals. Numbers were displayed one at a time, each with one of two nonsense labels; an example is DAP 364. Labels were said to indicate which of two groups the individual was a member of. After viewing 48 stimuli from each group, randomly mixed, the participant named the average for each group. Although the distributions were positively skewed, Malmi and Samson (1983) did not find a bias in estimates of the average of numbers, although Spencer (1963) did. Discrepancies between these studies using skewed distributions suggest that one cannot easily generalize results from one stimulus material to another, and that methods of presentation and response matter.

Several hypotheses can be tested with our skewed distributions. First, if participants use the most likely duration as an estimate of the mean, estimates will be biased toward the peak of a skewed distribution. Second, subjective estimates of probabilities are biased, with small probabilities overestimated and high probabilities underestimated (e.g., Luce & Suppes, 1965). If such subjective probabilities are used to weight stimulus durations, estimates of averages would be biased toward the long tail of a skewed distribution, whether that is positive or negative. Third, the contribution of a stimulus to the estimated average might be weighted by accumulated stimulus intensity (see Matthews, Stewart & Wearden, 2011, for recent discussion). If so, estimates would be biased toward long durations.

Finally, at the point when the participant is judging the average duration of a presented sample, the durations are in memory. After a delay, representations of durations in memory are sometimes shorter than the presented durations (e.g., Meck, 1983; Spetch & Wilkie, 1983). Estimates of means based on such subjectively shortened representations would underestimate population means. Subjective shortening follows from Scalar Timing Theory (Gibbon, 1977, 1981; Gibbon, Church & Meck, 1984; Meck, 1983), and leads to two simple but specific predictions. According to the theory, when a duration T is presented, a representation of it in long term memory has duration $D = BT + A$. The theory predicts, then, that if participants estimate the mean of presented durations as their mean duration in memory, the estimate would be a linear function of the mean presented duration. For our stimuli the theory also specifically predicts that variances of estimated means would be a linear function of presented means squared; details are below in the general discussion.

Wearden, Parry and Stamp (2002) propose that subjective shortening in memory only occurs with durations. If so, averaging of durations may be different from averaging other stimulus dimensions. However, shortening is sometimes reported for other dimensions. For

example, Moyer, Bradley, Sorensen and Whiting (1978) report a smaller Stevens' Law exponent for sizes in memory than for sizes perceived, resulting in shortening. Shortening of distances in head-mounted displays has also been reported (see Zhang, Nordman, Walker & Kuhl, 2012). Brown et al. (2005) say identification and discrimination of durations follow the same principles as for other stimulus properties. Although the proposition that durations are special is debatable, one cannot with certainty extrapolate to duration from results about other stimulus properties.

Experiment 1

The purpose of the first two experiments was to try out and refine a paradigm in which participants would be likely to estimate average durations reasonably well if they were able to do so at all; in particular, to see whether estimated averages of the three distributions with different means would differ. In Experiment 1, participants were first presented with a set of 40 tones with various durations sampled from one of three tone durations (symmetric, positively skewed, and negatively skewed), and were instructed to estimate the average of the 40 tone durations. Then, participants were presented with comparison tones, for each of which they had to decide whether it was longer than the average of the 40 sample tones. A within participants design is more sensitive for detecting differences than a between participants design, so in the first experiment each participant was presented with all three distributions, the distributions being varied between blocks of trials.

A drawback to a within participants design is that a response in one condition might be influenced by stimuli from other irrelevant conditions. It was initially not clear how demarcated a relevant series of tone durations would need to be to constrain this possibility. There is evidence that when people judge a recently presented duration they are influenced by previously presented durations (e.g., Jones & Wearden, 2004; Taatgen & van Rijn, 2011). On the other

hand, estimates of average are not always influenced by irrelevant stimuli. Malmi and Samson (1983) found that participants were able to separately estimate averages of two sets of stimuli randomly intermixed in a sequence. In a visual averaging task, Morgan et al (2000) found that when asked to do so, participants could make a judgment about the average of a specific sample of stimuli, ignoring other stimuli. This confirmed previous results by Morgan (1992), "Morgan's [1992] results indicate that the observer can *select* the stimuli that are relevant to a particular judgement, and ignore the others," (Morgan et al., 2000, p. 2345, emphasis in original). Further, Chong and Treisman (2005) found that when blue circles and green circles were randomly intermixed in the same display, participants were able to successfully estimate the average size of circles for each color separately. We anticipated that if participants were able to estimate the average of a set of durations, they would be able to base a judgment on the most recently presented block of tones, if instructed to do so.

Method

Participants. Fifteen undergraduate students at Purdue University were recruited from the subject pool of the introductory psychology course. They received course credits for participation. All participants reported having no hearing impairment. One participant's accuracy was very low; responses suggested the key meant to indicate "longer" was used to indicate "shorter," so the data from this participant were dropped.

Stimuli, task, and procedure. Experiments reported here all used the same three distributions of durations. We constructed a positively skewed distribution that is somewhat realistic for human activity times by modifying a distribution used to model human reaction times. Reaction times in many tasks can be approximated well by an ex-Gaussian distribution, the sum of a normal random variable and an independent exponential random variable (e.g., Hockley, 1984). The distribution has three parameters, μ and σ, the mean and standard deviation

of the normal distribution, and τ, the mean of the exponential distribution. We started with the parameter values of σ and τ that Hockley (1984) used to model the reaction time distribution in a memory search task with set size 4, when the probe was absent from the memory set (see Figure 4 in Hockley, 1984). These parameters were multiplied by three to produce durations in a range relatively easy to judge. The value μ = 300 ms was then selected to make the skew appreciable and to make the means appreciably different for our three distributions (symmetric, positively skewed, negatively skewed). The result was an ex-Gaussian positively skewed distribution, with μ = 300 ms and σ = 180 ms for its normal distribution, and τ = 662.1 ms for the mean of its exponential distribution. The symmetric distribution was normal, with mean 1535.7 ms and standard deviation $\sqrt{\sigma^2 + \tau^2}$ ms = 686.1 ms. Distributions were truncated, so values were between 3 and 3071.4 ms. The negatively skewed distribution was a mirror reflection of the positively skewed distribution. To generate samples from it, a sample from the untruncated positively skewed distribution was generated, each sample value was subtracted from 3071.4 ms, and any resulting values out of the range 3 to 3071.4 ms were replaced. Samples from all three distributions have approximately the same standard deviation and range. The positively skewed distribution is similar in shape to a typical reaction time distribution, but with a larger mean and a more pronounced degree of skewness. The random-number generator for the normal distribution was constructed with the Ziggurat method (see Marsaglia & Tsang, 2000). The three distributions resulting from 10,000 simulation trials for each distribution are shown in Figure 1.

The apparatus consisted of a personal computer and a 17-in. flat monitor. The experiment was controlled by E-Prime 1.1 (Psychology Software Tools). Participants were tested individually in a quiet room. Participants were seated in front of the screen at an unrestricted viewing distance of approximately 50 cm and wore headphones. They placed their right index,

middle, and ring fingers on the left, middle, and right keys, respectively, of a 5-key response box (Psychology Software Tools). The response box was to the right of the computer screen. The experimenter started the program, and participants read the instructions on the screen.

There were three phases in the experiment: familiarization, practice, and test. The objective of the first phase was to familiarize participants with comparison of durations. It consisted of five blocks including three trials. In each block, participants were presented with a sample tone followed by a comparison tone and asked to judge whether the comparison tone was longer or shorter than the sample tone. Tones were presented without graded onset or offset. Under our conditions, onsets and offsets were abrupt, with no noticeable clicks. Participants pressed the left and right keys to respond "shorter" and "longer", respectively. On each trial, presentation of the sample tone and the comparison tone was self-paced. Participants pressed the middle key when the word “Ready?” appeared on the screen, and a sample tone was presented immediately binaurally. After 1500 ms, the prompt "Comparison Ready?" appeared. When the participant pressed the middle key, the comparison tone was presented without delay. As soon as the comparison tone ended the message “Is this longer than the sample?” prompted participants to respond. Immediately after the response key was pressed, feedback was on screen for 1000 ms.

The duration of a sample tone was randomly drawn from a uniform distribution with range between 500 ms and 2000 ms, and the frequency was randomly chosen from 400, 450, 500, 550, 850, 900, 950, and 1000 Hz, one frequency in each block. The three comparison durations used in a block were 500, 1000, and 1500 ms, presented in random order. The frequency of the comparison tones was identical to that of the corresponding sample tone. If a response was correct, "CORRECT” was displayed on the screen; if it was an error, “ERROR” was displayed. The familiarization phase took less than three minutes.

In the practice phase participants practiced comparing their estimated average of a series (a block) of sample tone durations with a comparison tone duration. It was emphasized that they should not count to estimate time. Participants were told, "After all sample tones are presented, you will be asked to indicate whether a test tone duration is longer than the average of sample tones that you have just listened to." Participants were told that each block had a unique average duration, but it was not mentioned that there were three different distributions of tone durations.

At the start of a block, the prompt "Ready?" appeared. The participant pressed the middle key and 1000 ms later the first of a series of 20 sample tones with various durations was presented, with 1500 ms between them. The prompt "Comparison Ready?" appeared 1500 ms after the last sample tone ended. The participant pressed the middle key, and 1000 ms later the first of 20 comparison tones was presented. Immediately after a comparison tone ended, the prompt "Is this longer than the average?" appeared. Responses were as in the familiarization phase; that is, participants pressed the left key to indicate shorter and the right key to indicate longer. Comparison trials were self-paced as in the familiarization phase. Bauer (2009a) found that estimates of average are influenced by feedback. No feedback was given on accuracy of the participants' judgments.

Tone durations were randomly sampled from one of three distributions (symmetric, positively skewed, or negatively skewed) at the beginning of each block. The frequency of sample tones was 700 Hz. The display was blank during presentation of sample tones. Ten tone durations were used for the comparisons, 300, 600, 900, . . . , 3000 ms. Comparison tone durations were presented at random, with the constraint that each tone duration was presented twice. This produced 20 comparison trials in each block. The practice phase had three blocks, one for each distribution. It took approximately eight minutes to complete the practice phase.

The test phase immediately followed the practice phase. The test phase consisted of 12 blocks of trials. Each of the three distributions was presented in four blocks. The procedure was essentially identical to that of the practice phase, except that there were 50 sample tones in a block and the sample tone frequency was either 400, 500, 900, or 1000 Hz. Each frequency was used in one block for each distribution, the same frequency for sample tones and comparison tones. The order of the tone frequencies was randomized across the blocks as was the order of the distributions; a different randomization was done for each participant. After the sixth block, participants were allowed to take a break of up to five minutes. A participant's response ("shorter" or "longer") was recorded on each comparison trial. For each distribution for each comparison tone, over the test blocks there were eight observations for each participant; each participant's psychophysical function for a stimulus distribution is formed from 80 observations. The entire session lasted approximately an hour.

Results

Because each participant was presented with random samples of tone durations from each theoretical distribution, sample distributions were slightly different for each participant. For each participant, the mean, variance and skewness of the sample stimulus distributions were calculated. Table 1 shows the average of these over participants.

A participant's response to a comparison tone indicated whether it was judged longer than the average duration of the stimulus sample just presented. Proportions of "longer" responses were computed for each participant at each comparison duration to obtain a psychophysical function for each of the three distributions. Figure 2 shows the average psychophysical function over participants. The three stimulus distributions are clearly separated. From each participant's psychophysical function for each stimulus distribution, the distribution of that participant's

estimate of the average of the stimulus durations was estimated using the Spearman-Kärber method (Spearman, 1908). Miller and Ulrich (2001) provide a review of the method with calculation procedures. From each participant's estimated distribution, the mean, variance and skewness were calculated. Table 1 gives the average of these over participants.

The means of the three stimulus distributions differ. One test of whether participants are sensitive to the different distributions is to test whether their estimated averages of the three distributions differ. Each participant's estimated average was entered into a repeated measures Analysis of Variance (ANOVA) with Distribution as the factor. Distribution had a significant effect, $F(2, 26) = 55.03$, $MSE = 25587.60$, $p < .001$. We turn to comparing the means, variances and skewnesses of the presented distributions with those of the response distributions.

Mean. It is reasonable to assume means approximately meet the assumptions underlying t-tests, so paired t-tests are used to compare the stimulus and response means. For the symmetric stimulus distribution, the participants' estimate of the mean is unbiased; for the difference between the stimulus and response mean, $t(13) = 1.04$, n.s. But for the positively skewed stimulus distribution, the participants' estimate of the mean is positively biased, $t(13) = 4.99$, $p < .001$. And for the negatively skewed stimulus distribution, the participants' estimate of the mean is negatively biased, $t(13) = -4.20$, $p < .01$. There is a bias only for the skewed distributions, and it is in the direction of the long tail.

Variance. If participants are able to follow the instructions and respond based on their estimates of the mean, then each subject's response distribution is a sample from the sampling distribution of the mean. For a sample of size N, the variance of the sampling distribution of the mean is V/N, where V is the variance of the parent distribution (e.g., Hays, 1994, p. 213-215). Qualitatively, this predicts that the response variance would be smaller than the stimulus variance. However, suppose a stimulus of duration D produces a subjective impression of $D + e$,

where *e* is a random variable, an error caused by internal noise. If *e* has expected value 0, the average of (*D* + *e*) would be an unbiased estimate of the expected value of *D*. However, if the variance of *e* is large, the variance of the average of a sample of (*D* + *e*) could be larger than *V*[*D*]. With a large effect of internal noise, even an unbiased estimator of the population duration could have variance larger than the population variance.

For testing, the ratio of two sample variances would have an *F* distribution if they were calculated from two independent normally distributed random variables, but our situation is far from that. It is difficult to judge how well the distributions of response variances meet the assumptions underlying an *F*-test, even approximately. Bootstrap confidence intervals are suitable when underlying distributions are unknown (Efron & Tibshirani, 1993), so we compared variances with them. They were computed in R with procedures boot and boot.ci in the package boot. The following computations were carried out. To compare the stimulus and response variances, the difference between the variances of the stimulus distribution and the response distribution was calculated for each of the 14 subjects. Then 5000 random samples of size 14 were taken with replacement from the distribution of differences. For each sample of differences, the average difference was calculated. A 95% and a 99% confidence interval for the average difference were then formed from these 5000 bootstrap samples. Two-sided significance tests were based on these intervals. When a significant difference is reported, the larger of the two intervals that does not include 0 is reported.

For all three stimulus distributions, the variance of the response distribution is significantly smaller than that of the stimulus distribution. A 99% confidence interval for the response distribution variance minus the stimulus distribution variance for the symmetric distribution is [-262814 ms^2, -147327 ms^2], for the positively skewed distribution is [-254995

ms^2, -149274 ms^2], and for the negatively skewed distribution is [-228020 ms^2, -12391 ms^2]. Variances of response distributions, as expected, are smaller than those of stimulus distributions.

Skewness. Response distribution skewness is not tidy. For every participant every stimulus sample from the positively skewed parent distribution had positive skew, and every stimulus sample from the negatively skewed parent distribution had negative skew. For skewed stimulus distributions, the average skew of response distributions had the same sign as the skew of the stimulus distribution (Table 1). However, response distributions for some participants did not have skew of the same sign as that of the corresponding skewed stimulus distribution. Three of the 14 participants had negatively skewed response distributions for the positively skewed stimulus distribution and nine, a majority, had positively skewed response distributions for the negatively skewed stimulus distribution. For the symmetric distribution nine of 14 participants had positively skewed response distributions. Most response distributions were positively skewed, regardless of the stimulus distribution's skew. Skew of the stimulus distributions is not manifest well in that of the response distributions, and further analysis does not seem warranted.

Discussion

Results of Experiment 1 indicate that participants are able to produce different judgments of duration sample averages for samples drawn from populations with different means. Estimates of average duration are orderly, biased toward the long tail of the skewed distributions and unbiased for the symmetric distribution. There are ways the results could have occurred that would suggest participants considered only the relevant most recent block of durations when judging a comparison duration; for example, each response mean could have occurred at the peak of the corresponding stimulus mean. But the results that occurred are consistent with previously presented irrelevant blocks influencing participants. The mean of the three stimulus distributions combined is about equal to that of the symmetric distribution. The direction of the biases found

in Experiment 1 could be explained by participants using a weighted average of two means: the mean of the relevant block of durations and the mean of durations previously experienced during the experimental session. This possibility is addressed in the next experiment.

Experiment 2

Experiment 2 used a between participants design, each participant being presented with one distribution. Thus, the present procedure excluded the possibility that participants' estimates of average durations were influenced by tones in previous blocks that were sampled from irrelevant distributions.

Method

Participants. Twenty-five undergraduate students at Purdue University were recruited from the same subject pool as in Experiment 1. None had participated in Experiment 1. All reported having no hearing impairment. Participants were randomly assigned to the three distributions, eight to each. Inadvertently, an additional participant was assigned to the group presented with the symmetric distribution. One participant in the negatively skewed distribution condition had very low accuracy; responses suggested the key meant to indicate "longer" was used to indicate "shorter," so data from this participant were dropped.

Stimuli, task, and procedure. Apparatus, stimuli, and procedure were identical with those of Experiment 1, except for the following. Each participant was given eight test blocks with one of the three sampling distributions (symmetric, positively skewed, or negatively skewed). The tone frequency in a test block was one of 400, 450, 500, 550, 850, 900, 950 and 1000 Hz, each used once with the order random. The same frequency was used for the stimulus tones and the comparison tones in a block. There were three blocks of practice trials, as in Experiment 1, except that the same distribution was used in practice as in the test blocks to follow. Tone frequencies for practice blocks were 650, 700, and 750 Hz, each used once in random order.

Each participant's psychophysical function for a stimulus distribution was formed from 160 observations. An experimental session lasted approximately 40 minutes.

Results

Parameters were estimated in the same way as in Experiment 1, and are presented in Table 2. Psychophysical functions averaged over participants are in Figure 3. Response means and variances were analyzed as in Experiment 1.

Mean. Each participant's estimate of the mean was input to a one-way ANOVA with Distribution as the factor. There was a significant effect of Distribution, $F(2, 21) = 9.98$, $MSE = 40836.99$, $p < .001$. The estimated means differ for the three presented distributions.

For the symmetric stimulus distribution, although the participants' estimate of the mean is not significantly different from the stimulus mean, $t(8) = 2.16$, $p < .06$, the 95% confidence interval for the difference of means is [-8.78 ms, 273.76 ms], which barely excludes 0. For the positively skewed stimulus distribution, the participants' estimate of the mean is positively biased, $t(7) = 6.54$, $p < .001$. And for the negatively skewed stimulus distribution, the participants' estimate of the mean is negatively biased, $t(6) = -2.34$, $p < .05$.

Variance. Numerically, the variance of each response distribution is smaller than that of each stimulus distribution, as expected if response distributions are from the sampling distribution of the mean. However, the only stimulus distribution for which variance of the response distributions is significantly smaller than that of the stimulus distributions is the symmetric one. For it, a 95% confidence interval for response variance minus stimulus variance is [-206188 ms^2, -22091 ms^2]. For the positively and negatively skewed stimulus distributions, variance of the response distribution is not significantly different from that of the stimulus distribution. For the positively skewed stimulus distribution, a 95% confidence interval for the difference is [-154130 ms^2, 113638 ms^2]. For the negatively skewed distribution, a 95%

confidence interval for the response variance minus the stimulus variance is [-200549 ms^2, 24030 ms^2].

Skewness. Skewness of the positively skewed stimulus distribution was not well manifest in that of the response distribution; 7 of 9 participants had negatively skewed response distributions. For the other two stimulus distributions, stimulus and response skew were in better agreement. For the symmetric stimulus distribution, 4 of 8 had negatively skewed response distributions. For the negatively skewed stimulus distribution, 5 of 7 participants had negatively skewed response distributions.

Discussion

For the positively and negatively skewed stimulus distribution, participants' estimates of average are biased toward the long tail, the same result as in Experiment 1. The outcomes suggest that the bias is not due merely to the influences of tones from irrelevant distributions. But for the symmetric stimulus distribution the estimate of average is considerably larger than the stimulus distribution mean, unlike in Experiment 1.

A bias toward the upper tail of the symmetric distribution could be explained by a stronger influence of long durations than short ones, but this would lead to a bias toward the upper tail of each distribution, not found for the negatively skewed one. A more likely explanation for the directions of the biases is that participants based their estimates of averages on the sample durations combined with the comparison tone durations. In the bisection work mentioned earlier by Brown et al. (2005), the distribution of comparison tones had an effect on the subjective midpoint of two reference tones. In Experiment 2 here, the same distribution of comparison tone durations was used to test each stimulus distribution. Comparison tone durations ranged from 300 ms to 3000 ms, with mean 1650 ms. This is larger than the mean of the symmetric stimulus distribution, 1540 ms, and falls between the means of the positively and

negatively skewed stimulus distributions. Over the session if participants combined comparison tone durations with stimulus durations and based their estimates of average on the combination, the result would be estimates biased in the directions found.

For discussion, consider a participant half way through judging the 20 comparison durations for a single block. Fifty stimulus durations would have been presented, followed by 10 comparison durations. Suppose the current judgment is based on a mixture of these stimulus durations and comparison durations. (If the judgment is based on earlier blocks as well, reasoning is similar.) For such a mixture distribution, the mean is $(50m_D + 10m_C)/60$, where m_D and m_C are the means of the stimulus durations and comparison durations, respectively. For the positively skewed, symmetric, and negatively skewed conditions, the mixture distribution means are 1044, 1558, and 2073 ms, respectively. The first two are larger than the means of the corresponding stimulus distributions alone, the last smaller, the same pattern as was found in the response means.

A curious result in the response variances can be also explained with the mixture distributions. Numerically, the response variance in the positively skewed condition is larger than that in the symmetric condition. This is puzzling, because the stimulus variance in the positively skewed condition is not numerically larger than that in the symmetric condition. The variance of a mixture of 50 stimulus durations with 10 comparison durations is is $(50/60)s_D^2 + (10/60)s_C^2 + (50/60)(10/60)(m_D - m_C)^2$, where s_D^2 and s_C^2 are the variances of the stimulus and comparison durations, respectively (e.g., Townsend & Ashby, 1983, p. 264). For the positively skewed, symmetric and negatively skewed stimulus conditions, the variances of the mixture are 489,337; 444,051 and 444,758 ms^2, respectively. That for the positively skewed condition is larger than the other two, the same pattern as was found in the response variances. Results of Experiment 2 do not prove that participants combined comparison durations with stimulus

durations, but they are directed as one would expect if it were so. We note that comparison tone durations may have had an effect in Experiment 1, but their effect would not have been large because the larger number of irrelevant stimulus tone durations would dominate the biases. Experiment 3 was designed to avoid possible influence of irrelevant stimulus distributions and comparison tone durations.

Experiment 3

In Experiment 3, each comparison tone distribution was based on the sample tone distribution it tested, see below. As in Experiment 2, each participant was presented with durations sampled from only one stimulus distribution. Thus, the present procedure excluded the possible influences of tones from irrelevant distributions and reduced the possible influences from comparison tones. Aside from comparison tone durations, the method is the same as in Experiment 2.

Method

Participants. Forty-two undergraduate students at Purdue University were recruited from the same subject pool as used in the preceding experiments. No participant was in the previous experiments. All reported having no hearing impairment. Fourteen participants were randomly assigned to each of the three distributions.

Stimuli, task and procedure. Apparatus, stimuli, task, and procedure were identical with those of Experiment 2, except for the comparison tone durations. For each theoretical distribution, the durations at 10 percentiles were calculated; specifically, at percentiles 5, 15, 25, 35, 45, 55, 65, 75, 85 and 95. Those durations were used for the comparison tones. Each was used twice, making 20 presentations of comparison tones. Each participant's psychophysical function for a stimulus distribution was formed from 160 observations.

For the positively skewed distribution, the durations of the comparison tones were 218, 381, 496, 604, 719, 853, 1020, 1242, 1581, and 2308 ms. For the symmetric distribution, comparison tone durations were 407, 825, 1073, 1271, 1449, 1622, 1800, 1998, 2247, and 2664 ms. For the negatively skewed distribution, comparison tone durations were 763, 1490, 1829, 2051, 2218, 2352, 2467, 2575, 2690, and 2853 ms. We note that each comparison duration was presented twice. Each would be better matched to its corresponding stimulus duration if its number of presentations followed that stimulus distribution. But with such matching there would be different numbers of observations for different comparison durations, an undesirable cost.

Results

Parameter values, obtained as in the previous experiments, are in Table 3. Psychophysical functions averaged over participants are in Figure 4.

Mean. Means of the three response distributions are significantly different. A one-way ANOVA was conducted with each participant's response mean as the input and Distribution as the factor. There was a significant effect of Distribution, $F(2, 39) = 68.32$, $MSE = 39211.46$, $p < .001$.

For the symmetric stimulus distribution the participants' estimate is negatively biased; with a paired t-test the participants' estimate of the mean is significantly less than the stimulus mean, $t(13) = -2.78$, $p < .05$. For the positively skewed stimulus distribution, the participants' estimate of the mean is unbiased, $t(13) = .89$, n. s.. And for the negatively skewed stimulus distribution, the participants' estimate of the mean is negatively biased, $t(13) = -5.06$, $p < .001$.

As in the previous experiments, for the negatively skewed stimulus distribution, the participants' estimate of the mean is biased, toward the long tail. But in contrast with the previous experiments, for the positively skewed distribution the participants' estimate of the mean

is unbiased. Also differing from the previous experiments, the participants' estimate of the mean of the symmetric distribution is biased, toward the left tail.

Variance. Variance of the response distribution is smaller than that of the stimulus distribution for the symmetric and positively skewed distributions. A 99% confidence interval for the difference is [-258284 ms^2, -110211 ms^2] for the symmetric distribution and [-294227 ms^2, -177129 ms^2] for the positively skewed distribution. For the negatively skewed stimulus distribution however, there is no significant difference between the variances; a 95% confidence interval for the difference is [-99868 ms^2, 200070 ms^2].

Skewness. Skewness of the response distributions was in good, but not perfect, agreement with that of the stimulus distributions. Combining the skewed stimulus distributions, response skewness and stimulus have the same sign for 21 of 28 participants, significant at the .05 level with a sign test.

Comparing Experiments 2 and 3. The only difference in the designs of Experiments 2 and 3 is in the comparison durations. These were the same for all stimulus distributions in Experiment 2 but selected to be percentiles of the stimulus distributions in Experiment 3. This difference in comparison duration selection lead to differences in response means for the two experiments; specifically, for the symmetric and positively skewed distributions, response means of Experiment 3 were significantly different from those of Experiment 2. Each stimulus distribution was tested with an unpaired *t*, assuming unequal variances, conducted on response means of Experiments 2 and 3. For the symmetric distribution, $t(15) = 3.11$, $p < .01$; for the positively skewed distribution, $t(11) = 5.64$, $p < .001$, and for the negatively skewed distribution, $t(18) = 1.46$, n. s. The response mean differences show that the choice of comparison durations matters.

Scalar Timing Theory Predictions. A striking pattern in the results is that the response mean increases linearly with the stimulus mean, $R^2 = .999$ (see Figure 5), in accord with Scalar Timing Theory (Gibbon, 1977). Further, as shown in Figure 6, although the theoretical stimulus distributions all have the same variance, the response variance increases nearly linearly with the stimulus mean squared, $R^2 = .970$. (For response variance as a linear function of stimulus mean, $R^2 = .919$). These results are in accord with Scalar Timing Theory; derivations are in the General Discussion.

Discussion

In Experiment 3, participants were presented with samples of tone durations from a single stimulus distribution. They then judged whether comparison tone durations were larger than the average duration of the sample. Comparison tone durations were chosen to be evenly spaced percentiles of the stimulus distribution. The circumstances were designed to minimize biases from irrelevant stimulus distributions and the distribution of comparison tones. Under these circumstances, participants' estimates of average stimulus durations were unbiased for the small stimulus mean, but underestimated the medium and large stimulus mean, underestimation larger for the larger mean. For the durations investigated in Experiment 3, response means increased linearly with stimulus means, and response variances increased nearly linearly with stimulus means squared. A linear relation between response variance and the square of mean stimulus duration is at the core of Scalar Timing Theory (Gibbon, 1977).

General Discussion

The three distributions of durations employed here are a small selection from those one could consider. Nonetheless results with them eliminate several mechanisms as sole sources of the bias. Participants might estimate central tendency with the median or mode (peak) rather than

the mean. But for our negatively skewed distribution, the mean, median and mode occur in that order. Estimates of the mean for this distribution were significantly smaller than the stimulus mean in every experiment, and hence smaller than the stimulus median and mode. These can be rejected as estimates of the mean.

Subjective estimates of individual stimulus durations might be biased, thus producing a biased estimate of the mean. For example, a physical duration might be transformed to a subjective estimate in accordance with Stevens' Law, or the subjective estimate of a tone duration might depend on intensity accumulated over time, greater for long tones than short ones. The way biases of individual durations combine to produce a biased estimate of the mean depends on the details. There are two broad ways to consider the processing of the stimulus and the comparison tones. One is to assume that stimulus and comparison durations are transformed the same way. The other is to assume that they are not transformed the same way, because at the time of judgment a comparison tone is in sensory or working memory, but the stimulus tones are represented in long term memory.

Suppose the same transformation is applied to the stimulus durations and the comparison durations. The hypothesis that long durations are weighted more heavily than short ones, due to greater accumulated intensity, is immediately eliminated by the underestimation of means found in Experiment 3. One can also immediately eliminate the hypothesis that biased mean estimation is due to biased subjective estimates of probabilities, with small probabilities overestimated and high probabilities underestimated (e.g., Luce & Suppes, 1965). This mechanism cannot alone account for the biased estimate of mean for the symmetric distribution found in Experiment 3.

A general approach can be based on a Taylor expansion if the transformation has one. How would a comparison duration of exactly the mean of the stimulus durations be judged

relative to the mean of the transformed durations? Let $f(x)$ be the transformed value of physical duration x. Then a Taylor expansion about the mean stimulus value $\bar{x}$ is

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + f''(\bar{x})(x - \bar{x})^2/2 + f'''(\bar{x})(x - \bar{x})^3/6 + \ldots$$

The average transformed value of n presented durations $x_1, \ldots, x_n$ is

$$\overline{f(x)} = \Sigma f(\bar{x})/n + f'(\bar{x})\Sigma(x_i - \bar{x})/n + f''(\bar{x})\Sigma(x_i - \bar{x})^2/2n + f'''(\bar{x})\Sigma(x_i - \bar{x})^3/6n + \ldots$$

$$= f(\bar{x}) + f''(\bar{x})S^2/2 + f'''(\bar{x})M_3/6 + \ldots,$$

where S^2 is the variance and M_3 is the third moment about the mean. Whether the bias is positive or negative depends on the sign of the derivatives and of the moments.

With Stevens' Law, $f(x) = ax^b$, with $a, b > 0$. For averaging of brightness, Bauer (2009b) found that "to a first approximation, the perceived average . . . follows the same power function as for . . . a single object." This can be explained by the Taylor expansion, in which the first term is Stevens' Law applied to the mean, resulting in an approximation of $a\bar{x}^b$. For durations of white noise stimuli, the exponent is reported as $b = 1.1$ in Stevens (1975). For such a value of b between 1 and 2, the second derivative, $ab(b-1)\bar{x}^{b-2}$, is positive, so the second term is positive. For a normal distribution, odd numbered moments are 0 and even numbered moments are positive. The coefficient of an even numbered moment, e.g., $ab(b-1)(b-2)(b-3)\bar{x}^{b-4}/4!$ is also positive. Hence, for the normal stimulus distribution used in Experiment 3 the result is an overestimation of the stimulus mean, contrary to what was found. In an extensive review of power function exponents for duration, Eisler (1976) reports a wide range of values of b, from less than .5 to over 1.2, with average approximately .9. Values of b less than 1 are capable of predicting overestimations of small stimulus means and underestimations of large means, depending on the values of a and b. But it is difficult to properly pursue this further without knowing values of a and b for our stimulus distributions.

In a recent model of Taatgen, van Rijn and Anderson (2007) stimulus intervals are timed with an internal clock that produces ticks with intervals between them that are not constant but tend to increase over time. The first tick after time 0 occurs at time t_0. Let t_n denote the interval between tick n and tick $n + 1$. Then the interval between tick $n + 1$ and tick $n + 2$ is $t_{n+1} = at_n + e_n$, where e_n is a random variable with a logistic distribution that has mean 0 and standard deviation bat_n, and a and b are constants. Parameter estimates are $t_0 = 100$ ms, $a = 1.02$, and $b = .015$ (van Rijn & Taatgen, 2008, p. 369, updated from Taatgen et al., 2007, p. 582). When a stimulus interval to be timed is presented, the internal clock is set to time 0 at the start of the stimulus interval. Then ticks occur at times t_o, $t_o + t_1$, $t_o + t_1 + t_2$, . . . When the stimulus interval ends, the time of the last tick to precede the end is taken as the subjective duration of the stimulus interval (see Taatgen, et al., 2007, Figure 4), and is stored in memory.

This internal clock module was embedded in the pool model of the ACT-R cognitive architecture (Anderson, 2007; Anderson, et al., 2004). With this architecture, when a series of stimulus durations is presented, their mean as estimated by the subject can be predicted as weighted average of the subjective durations stored in memory. Formulas for the weighted average are in Taatgen and van Rijn (2011, Equations 1, 2, 3), given here in the Appendix.

To see what the timing and weighted averaging components of the model predict for our three stimulus distributions, a simulation[1] was run in R. Parameter d was set to the default value of .5. In the simulation, 12,000 values were randomly generated from each stimulus distribution (with positive, symmetric, or negative skew) to make a pool for that distribution. Then values less than 3 ms or larger than 3071.4 ms were deleted from the pool.

In the simulation, each participant was presented with 400 stimulus durations because in Experiment 3 each participant was presented with 8 test blocks of 50 stimulus durations, 400 total. In the model, weights for the weighted average depend on the time that elapsed between

the presentation of each stimulus tone and the time at which the weighted average is calculated. But these times are unknown, because the participant started the comparison tone phase of a block by pressing a key when ready, and presentation of individual comparison tones was self-paced. The longer the elapsed times, the more influence the weights have, so to give the weights a good deal of influence, elapsed times were measured from stimulus presentation to the end of the entire session, and the unknown start and end times of each block were ignored. A session of Experiment 3 for a participant took approximately 40 minutes. In the simulation, each participant was presented with 400 stimulus durations, spaced over all of 40 minutes except for a 5 second pause after the last stimulus presentation. For each stimulus distribution, 100 participants each received 400 durations from the corresponding distribution pool, randomly sampled without replacement. Presentation times of the 400 durations were evenly spaced from 1 sec to 2395 sec inclusively. For each participant, each duration was timed with the internal clock, producing a subjective duration. Then the weighted average of subjective durations was formed, under the assumption that the time at which subjective durations are retrieved, the current time, is 2400 sec. The weighted average for a participant is the estimated mean of the subjective distributions for that participant. The mean, variance and skewness of the estimated means for the simulated 100 participants were then calculated. Simulation of 100 participants was repeated 1000 times, using the same stimulus distribution pools each time. Means over the stimulus distribution pools and 1000 repetitions of 100 participants are in Table 4.

Some aspects of our Experiment 3 data are found in simulated results of the model. Specifically, simulated estimated means are smaller than stimulus means for symmetric and negatively skewed stimulus durations. Simulated estimated means increase linearly with the stimulus means (R^2= .99998). Furthermore, simulated response variances for the positive and symmetric distributions are smaller than the corresponding stimulus variances, as in the data. But

in more detail, the slope of the regression line predicting simulated estimated means from stimulus means is 1.00; that is, estimated means differ from the stimulus means by a constant, unlike the data. Also, variances produced by the simulations are not monotonically increasing with means. Further, variances of estimated means are only moderately well predicted as a linearly increasing function of stimulus means squared; the slope of the regression line is .01 and R^2 = .57. The simulations omit some details of the experimental paradigm. Nonetheless they indicate that the timing and weighted averaging components of the model, when applied to the stimulus distributions, produce some aspects of the data but not all.

Subjective Shortening in Memory

A source of bias leading to underestimation of individual durations is subjective shortening in memory. At the time participants make judgments about the average presented duration, the presented durations are represented in long term memory. Previous work shows that when durations are tested after a delay, durations in memory are sometimes shorter than the presented durations (e.g., Meck, 1983; Spetch & Wilkie, 1983). Some say subjective shortening in memory only occurs with durations as stimuli, and not, for example, with line lengths (e.g., Wearden, Parry & Stamp, 2002).

Subjective shortening of individual durations in memory, with long durations shortened more than short ones, would explain the underestimation of the medium and large stimulus means in Experiment 3. The estimated mean of the symmetric distribution is negatively biased and that of the negatively skewed distribution is more so. An objection to this explanation is that a significant bias did not occur for the positively skewed distribution. However, a 95 % confidence interval is [– 57 ms, 138 ms]. The lower bound is consistent with a small albeit nonsignificant negative bias.

Model. The following simple model accounts well for the data. In particular, it explains the two relations found in Experiment 3, (a) response means are linear functions of stimulus means and (b) variances of response means are linear functions of stimulus means squared. The model is for durations represented in memory (Gibbon, 1981; Gibbon, Church & Meck, 1984; Meck, 1983). It has been applied to bisection experiments, in which an organism indicates which of the two reference durations a comparison duration is more similar to. According to the model, after a duration T is presented its later representation D in reference (long term) memory has duration $D = BT + A$, where A and B are random variables. The transformation is produced by the process that stores the duration. Briefly, the speed of storage is Y, and the duration T in working memory is stored as $D = A + T/Y = BT + A$, where $B = 1/Y$ (Meck, 1983). When a comparison duration is presented it is in working memory while it is compared with a reference memory representation; in other words, the transformation is not applied to comparison durations.

We assume that the mean of a number of durations represented in memory is estimated simply as their average. To reduce the number of parameters, suppose A is constant. Let $E[X]$ and $V[X]$, respectively, denote the expected value and variance of a random variable X, and let $\overline{X}$ denote the mean of a sample of observations of X. If B and T are independent, then $E[D] = E[BT] + A = E[B]E[T] + A$. Then because $E[\overline{D}] = E[D]$ and $E[\overline{T}] = E[T]$ the average of a sample of D values is predicted to be a linear function of the average of the corresponding T values. This is relation (a).

The variance of D is more complicated. The constant A contributes nothing to the variance. For B and T independent, the variance of their product is

$$V[D] = V[BT] = E[B]^2V[T] + V[B]V[T] + V[B]E[T]^2$$

(Goodman, 1960; Gibbon, 1981, p. 81). For a sample of N independent observations,

$$V[\overline{D}] = V[D]/N = E[B]^2V[T]/N + V[B]V[T]/N + V[B]E[T]^2/N$$

$$= E[B]^2V[T]/N + V[B]V[T]/N + V[B]E[\overline{T}]^2/N.$$

For our stimuli, the population variance of presented durations, $V[T]$ is approximately the same for every presented distribution. The result is $V[\overline{D}]$ linear with $E[\overline{T}]^2$, relation (b).

To fit the model, for each of the three distributions, values of $E[T]$ and $V[T]$ in the equations above were estimated as the mean and variance for the stimulus distribution in Table 3. For each distribution, the response mean and variance in Table 3 were used as estimates of $E[\overline{D}]$ and $V[\overline{D}]$, respectively. It is not obvious what value to use for N, because as more blocks are presented, the number of durations that have been presented, i.e., the sample size, increases. Experiments 1 and 2 indicate that participants base their comparisons on more durations than the most recently presented block, but do not establish how far back the basis goes. If participants used only the most recent block of durations, the sample size N is 50. At the other extreme, each participant might use durations from all blocks. By the end of the experiment, each participant was presented with 400 durations, giving N of 400. For these two values of N the unknown parameters, $E[B]$, $V[B]$ and A were estimated to minimize the sum of squared errors using Solver in Excel. Response variances are much larger numerically than response means. If response variances were predicted, their squared errors would dominate the sum of squared errors, so response standard deviations were predicted.

For $N = 50$, parameter estimates were $A = 321.4$, $E[B] = .702$, $V[B] = 3.64$. For $N = 400$, parameter estimates were $A = 321.5$, $E[B] = .701$, and $V[B] = 29.51$. There are only six observations to predict, of course. Nonetheless, the proportion of variance in the observations accounted for by the model is remarkably high, $R^2 = .99881$ when $N = 50$ and $R^2 = .99878$ when

$N = 400$. Parameter $V[B]$ is sensitive to the value of N; the other parameters and goodness of fit are not. (Consequently, one cannot use the model to estimate the number of durations participants base their judgments on, by finding the best fitting value of N.)

At first the variance of B may seem large, given $B > 0$, with expected value about .70. But in the model, B is the reciprocal of a random variable Y. To check whether parameter values for B are reasonable, Y was assumed to be uniformly distributed between 0 and 10 and 1000 random values of Y were generated in Excel. For $1/Y$ the mean was .767 and the variance was 24.43. These are comparable to estimates for B with $N = 400$.

It is reasonable to consider reducing the number of parameters by omitting the intercept, A. When this is done, however, prediction errors are relatively large for the skewed distribution means; for each the observed mean deviates from the predicted mean in the direction of the long tail of the distribution.

Predicted and observed values are presented in Table 5, for the slightly worse fit with $N = 400$. The model, developed for similarity judgments with two reference durations (bisection), predicts well the mean and standard deviation of many subjective durations. Good predictions were obtained without considering skewness of the presented distributions.

Because skewness is a function of the third moment about the mean, we considered testing the prediction for this moment of the simple model $D = A + BT$. It is straightforward to derive an equation analogous to those above for $E[D]$ and $V[D]$. But testing it is infeasible. The equation has many terms, and, of more importance, estimates of the third moment about the mean are noisy. Estimates of response skewness are somewhat orderly. In each experiment they are monotonic with stimulus skew, from positive to symmetric to negative, although in Experiment 2 a sign of response skew does not agree with that of stimulus skew. Manipulation of skewness was part of the design of our experiments, but we find it difficult to pursue quantitatively.

Conclusion. Our experiments indicate that people are fairly good at estimating the average of a sample of durations. Experiments 1 and 2 show that estimates are influenced by previously presented irrelevant durations, as in earlier findings by, e.g., Jones and Wearden (2004) and Taatgen and van Rijn (2011). Estimates are also influenced by comparison durations used for testing, as in earlier findings by, e.g., Brown et al. (2005). Results are in accord with the finding by Jones and McAuley (2005) that distributional properties of global temporal context have effects on time judgments of sequences of isolated time intervals. Results are also in accord with previous findings of effects of contexts on magnitude judgment. According to Helson's (1964) Adaptation Level Theory, judgments are influenced by the global context during the experiment, more specifically by the current stimulus value, values of stimuli experienced prior to the current stimulus, and values of other relevant stimuli present during the current trial (see also Parducci, Calfee, Marshall, & Davidson, 1960).

Our results differ from the finding that with certain visual stimuli participants are able to estimate the average of some property, ignoring that property in irrelevant stimuli (Chong & Treisman, 2005; Morgan et al., 2000). With our tone durations relevant stimuli were indicated by pitch. With our set up, this was not an adequate cue for participants to use for basing their estimate of average duration on only durations in the most recent block. The objective of the present study was not to investigate these sources of influence on judgments of averaged durations, but rather to detect and remove them.

In Experiment 3 participants were presented with a single distribution of durations, and comparison durations were chosen to be evenly spaced percentiles of that duration. A simple single mechanism explains the main results of Experiment 3: subjective shortening of individual durations represented in reference memory, with more shortening for longer durations. Subjective shortening in memory is said by some to be unique to durations (e.g., Wearden, Parry

& Stamp, 2002). It would lead to special problems in estimating average durations of events experienced in everyday life. Estimates would typically be underestimations, producing frequent surprises when events take longer than expected.

Note

1. We thank Hedderik van Rijn for kindly sending us a program in R for simulations of the Taatgen, van Rijn and Anderson (2007) model. We modified the program slightly.

References

Albrecht, A. R., Scholl, B. J., & Chun, M. M. (in press). Perceptual averaging by eye and ear: Computing summary statistics from multimodal stimuli. *Attention, Perception & Psychophysics*. doi 10.3758/s13414-012-0293-0

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Science, 15,* 122-131. doi:10.1016/j.tics.2011.01.003

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M., Douglass, D., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review, 111,* 1036-1060.

Baker, C. H. (1962). On temporal extrapolation. *Canadian Journal of Psychology, 16,* 37-41. doi: 10.1037/h0083223

Bauer, B. (2009a). The danger of trial-by-trial knowledge of results in perceptual averaging. *Attention, Perception & Psychophysics, 71,* 655-665. doi:10.3758/APP.71.3.655

Bauer, B. (2009b). Does Stevens' Power Law for brightness extend to perceptual brightness averaging? *The Psychological Record, 59,* 171-186.

Brown, G. D. A., McCormack, T., Smith, M., & Stewart, N. (2005). Identification and bisection of temporal durations and tone frequencies: Common models for temporal and nontemporal stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 919-938. doi: 10.1037/0096-1523.31.5.919

Chong, S. C., & Treisman, A (2003). Representation of statistical properties. *Vision Research, 43,* 393-404. doi: 10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45,* 891-900. doi:10.1016/j.visres.2004.10.004

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

Eisler, H. (1976). Experiments on subjective duration 1868-1975: A collection of power function exponents. *Psychological Bulletin, 83,* 1154-1171. doi: 10.1037/0033-2909.83.6.1154

Gibbon, J. (1977). Scalar expectancy theory and Weber's Law in animal timing. *Psychological Review, 84,* 279-325. doi: 10.1037/0033-295X.84.3.279

Gibbon, J. (1981). On the form and location of the psychometric bisection function for time. *Journal of Mathematical Psychology, 24, 58-87.* doi: 10.1037/0097-7403.8.3.226

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. In J. Gibbon & L. Allan (Eds.), *Timing and time perception* (Vol. 423, pp. 52-77). NY: New York Academy of Sciences. doi:10.1111/j.1749-6632.1984.tb23417

Goodman, L. (1960). On the exact variance of products. *Journal of the American Statistical Association, 55,* 708-713. doi:10.1080/01621459.1960.10483369

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 718-734. doi: 10.1037/a0013899

Hays, W. L. (1994). *Statistics.* Fort Worth: Harcourt Brace.

Helson, H. (1964). *Adaptation-level theory*. Oxford: Harper & Row.

Hockley,W. E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 598-615. doi: 10.1037/0278-7393.10.4.598

Jarvstad, A., Rushton, S. K., Warren, P. A., & Hahn, U. (2012). Knowing when to move on: Cognitive and perceptual decisions in time. *Psychological Science, 23,* 589-597. doi: 10.1177/0956797611426579.

Jones, M. R., & McAuley, J. D. (2005). Time judgments in global temporal contexts. *Perception & Psychophysics, 67,* 398-417. doi: 10.3758/BF03193320

Jones, L. A., & Wearden, J. H. (2004). Double standards: Memory loading in temporal reference memory. *Quarterly Journal of Experimental Psychology, 57B,* 55-77. doi: 10.1080/02724990344000088

Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology: Vol. 3* (pp. 249-410). NY: Wiley.

Malmi, R. A., & Samson, D. J. (1983). Intuitive averaging of categorized numerical stimuli. *Journal of Verbal Learning and Verbal Behavior, 22,* 547-559. doi:10.1016/S0022-5371(83)90337-7

Marsaglia, G., & Tsang, W. W. (2000). The Ziggurat method for generating random variables. *Journal of Statistical Software, 5* (http://www.jstatsoft.org/v05/i08/paper). Retrieved December 3, 2006. doi:10.1145/355744.355749

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,* 303-313. doi: 10.1037/a0019961

Meck, W. H. (1983). Selective adjustment of the speed of internal clock and memory processes. *Journal of Experimental Psychology: Animal Behavior Processes, 9,* 171-201. doi: 10.1037/0097-7403.9.2.171

Miller, J., & Ulrich, R. (2001). On the analysis of psychometric functions: The Spearman-Kärber method. *Perception & Psychophysics, 63*, 1399-1429. doi: 10.3758/BF03194551

Morgan, M. J. (1992). On the scaling of size judgements by orientational cues. *Vision Research, 32*, 1433–1445. doi: 10.1016/0042-6989(92)90200-3

Morgan, M.J.,Watamaniuk, S.N.J., McKee, S.P. (2000). The use of an implicit standard for measuring discrimination thresholds. *Vision Research, 40*, 2341-2349. doi: 10.1016/S0042-6989(00)00093-6

Moyer, R. S., Bradley, D. R., Sorensen M. H., & Whiting, J. C. (1978), Psychophysical functions for perceived and remembered size. *Science, 200*, 330-332. doi: 10.1126/science.635592

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4,* 739-744. doi: 10.1038/89532

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*, 407-418. doi: 10.1037/h0022602

Parducci, A., Calfee, R. C., Marshall, L. M., & Davidson, L. P. (1960). Context effects in judgments: Adaptation level as a function of the mean, mid-point, and median of the stimuli. *Journal of Experimental Psychology, 60*, 65-77. doi: 10.1037/h0044449

Parducci, A., & Perrett, L. F. (1971). Category rating scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology, 89*, 427-452. doi: 10.1037/h0031258

Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. *Psychological Science,* published

online 12 June 2013. doi: 10.1177/0956797612473759

Spearman, C. (1908). The method of right and wrong cases (constant stimuli) without Gauss's formulae. *British Journal of Psychology, 3*, 227-242.

Spencer, J. (1961). Estimating averages. *Ergonomics, 4*, 317-328. doi: 10.1080/00140136108930533

Spencer, J. (1963). A further study of estimating averages. *Ergonomics, 6*, 255-265. doi:10.1080/00140136308930705

Spetch, M. L., & Wilkie, D. M. (1983). Subjective shortening: A model of pigeon's memory for event duration. *Journal of Experimental Psychology: Animal Behavior Processes, 9*, 14-30. doi: 10.1037/0097-7403.9.1.14

Stevens, S. S. (1975). *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New York: Wiley & Sons, 1975.

Taatgen, N. A., & van Rijn, H. (2011). Traces of times past: Representations of temporal intervals in memory. *Memory & Cognition, 39*, 1546-1560. doi: 10.3758/s13421-011-0113-0

Taatgen, N. A., van Rijn, & Anderson, J. (2007). An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning. *Psychological Review, 114*, 577-598. doi: 10.1037/0033-295X.114.3.577

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes.* Cambridge: Cambridge University Press.

Wearden, J. H., Parry, A., & Stamp, L. (2002). Is subjective shortening in human memory unique to time representations? *The Quarterly Journal of Experimental Psychology, 55B*, 1-25. doi: 10.1080/02724990143000108

Zhang, R., Nordman, A., Walker, J., & Kuhl, S. A. (2012). Minification affects verbal- and action-based distance judgments differently in head-mounted displays. *ACM Transactions on Applied Perception, 9,* 14:1-14:13. doi: 10.1145/2325722.2325727

Appendix A

Weighted Average for Simulation of Taatgen et al. (2007) Model

The following equations are from Taatgen and van Rijn (2011, Equations 1, 2, 3), with minor notation changes. They lead to the weighted average of a series of presented subjective durations, used in simulations here as the participant's estimated mean of the corresponding stimulus durations.

Let t_{current} be the time at which subjective duration V_j is retrieved from memory and let t_{creation} be the time at which it was created. The activation of subjective duration j is

$$A(t_{\text{current}}) = \log(t_{\text{current}} - t_{\text{creation}})^{-d} + \text{mismatchpenalty}.$$

In our experiments, mismatchpenalty is not relevant, set to 0.

The weight of subjective duration j is

$$P_j = [\exp (A_j(t_{\text{current}})/t)]/\Sigma_i[\exp (A_i(t_{\text{current}})/t)],$$

where t is a noise parameter, set to .2 (when t_{current} has units s).

The weighted average of the subjective durations $V_1, \ldots V_j, \ldots$ is

$$\Sigma_j P_j V_j.$$

Table 1

Experiment 1: Parameters of Stimulus and Response Distributions

	Positive Skew		Symmetric		Negative Skew	
	Stimuli	Responses	Stimuli	Responses	Stimuli	Responses
Mean (ms)	934	1216	1534	1594	2148	1846
	(43)	(216)	(41)	(222)	(67)	(262)
Variance (ms^2)	349,261	153,750	373,682	167,310	342,907	240,629
	(38,620)	(78,742)	(28,223)	(80,662)	(61,108)	(141,440)
Skewness	1.11	0.64	-0.00	0.21	-1.14	-0.21
	(0.17)	(0.93)	(0.11)	(0.68)	(0.16)	(0.85)

Note: Standard deviations in parentheses.

Table 2

Experiment 2: Parameters of Stimulus and Response Distributions

	Positive Skew		Symmetric		Negative Skew	
	Stimuli	Responses	Stimuli	Responses	Stimuli	Responses
Mean (ms)	923	1498	1540	1672	2157	1962
	(27)	(241)	(29)	(201)	(29)	(146)
Variance (ms^2)	342,833	334,796	374,525	259,969	334,595	245,872
	(38,251)	(217,411)	(16,211)	(140,821)	(35,610)	(155,818)
Skewness	1.12	-0.02	-0.01	-0.52	-1.14	-0.89
	(0.12)	(0.70)	(0.07)	(0.64)	(0.10)	(1.10)

Note: Standard deviations in parentheses.

Table 3

Experiment 3: Parameters of Stimulus and Response Distributions

	Positive Skew		Symmetric		Negative Skew	
	Stimuli	Responses	Stimuli	Responses	Stimuli	Responses
Mean (ms)	921	961	1549	1419	2168	1836
	(24)	(159)	(36)	(174)	(40)	(250)
Variance (ms^2)	333,683	102,808	352,454	173,474	338,229	389,134
	(38,946)	(81,867)	(22,727)	(107,607)	(32,185)	(321,910)
Skewness	1.36	0.57	0.00	-0.44	-1.20	-0.59
	(.12)	(.81)	(.08)	(1.00)	(.10)	(.96)

Note: Standard deviations in parentheses.

Table 4

Simulation of Taatgen et al. (2007) Model

	Positive Skew		Symmetric		Negative Skew	
	Stimuli	Estimates	Stimuli	Estimates	Stimuli	Estimates
Mean (ms)	924	862	1,540	1,474	2,149	2,089
Variance (ms^2)	339,568	223,057	393,840	281,512	341,784	276,182
Skewness	1.12	1.06	-.01	.14	-1.14	-.56

Table 5

Observed Response Means and Standard Deviations and their Predictions from the Scalar Timing Theory Model

	Positive Skew		Symmetric		Negative Skew	
	Observed	Predicted	Observed	Predicted	Observed	Predicted
Mean (ms)	961	967	1419	1408	1836	1842
Standard Deviation (ms)	321	296	417	451	624	610

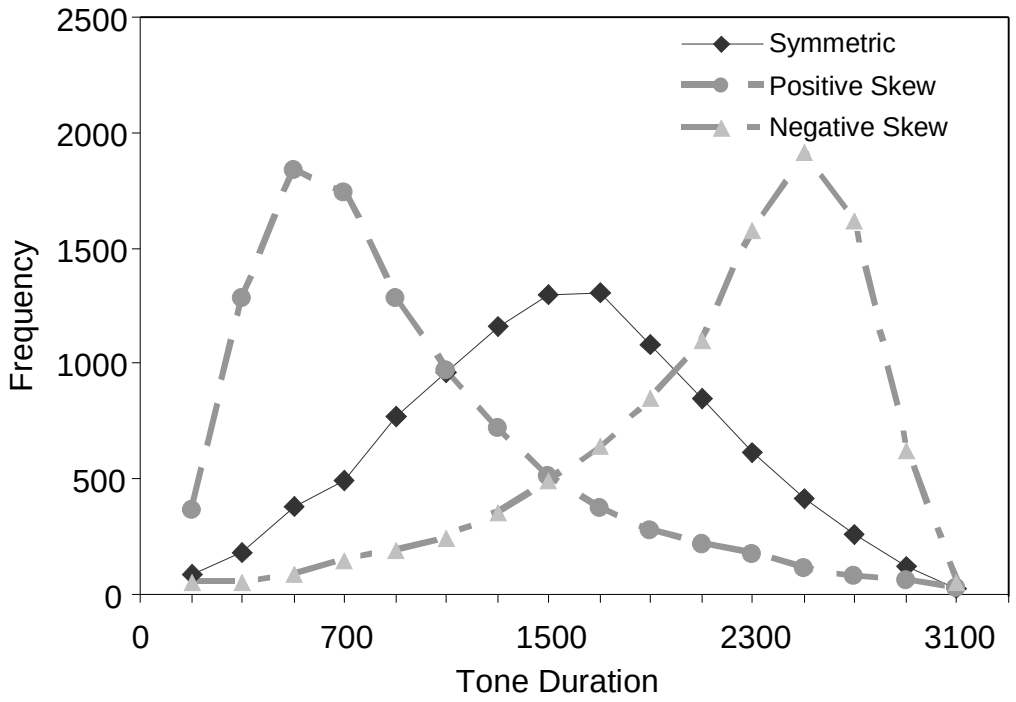

Figure 1. The three distributions of tone durations (in ms) used in the present study. Each illustrated distribution is based on 10,000 samples.

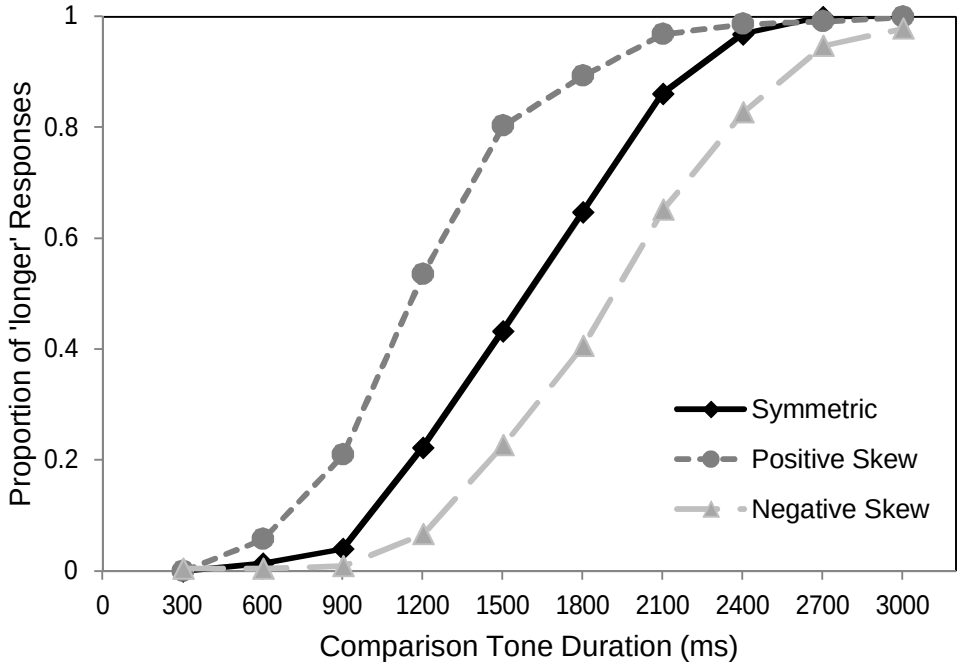

Figure 2. Psychophysical functions averaged over all participants, Experiment 1.

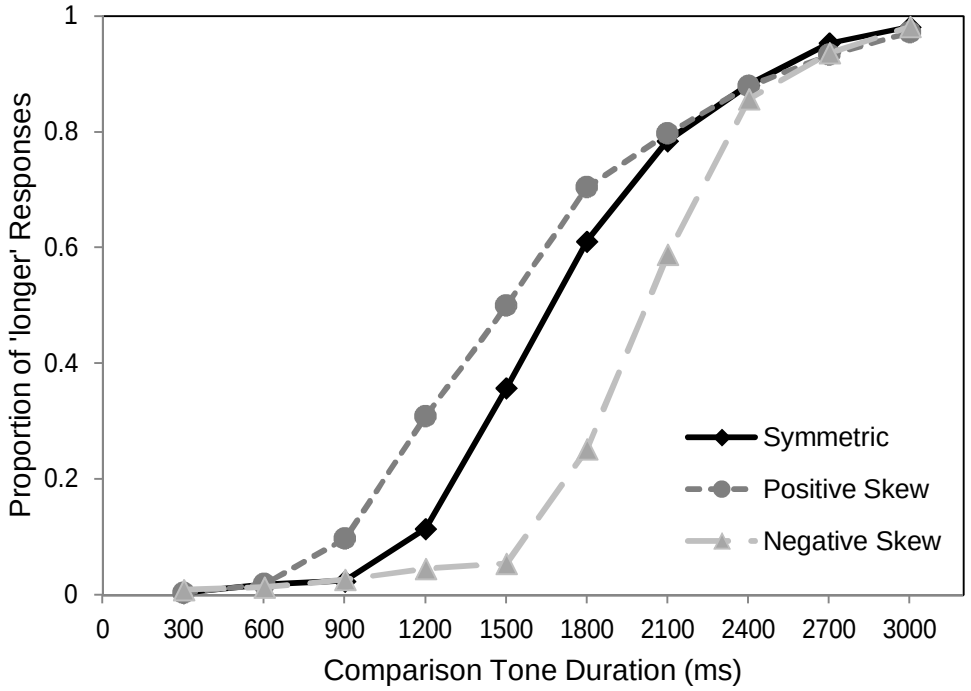

Figure 3. Psychophysical functions averaged over all participants, Experiment 2.

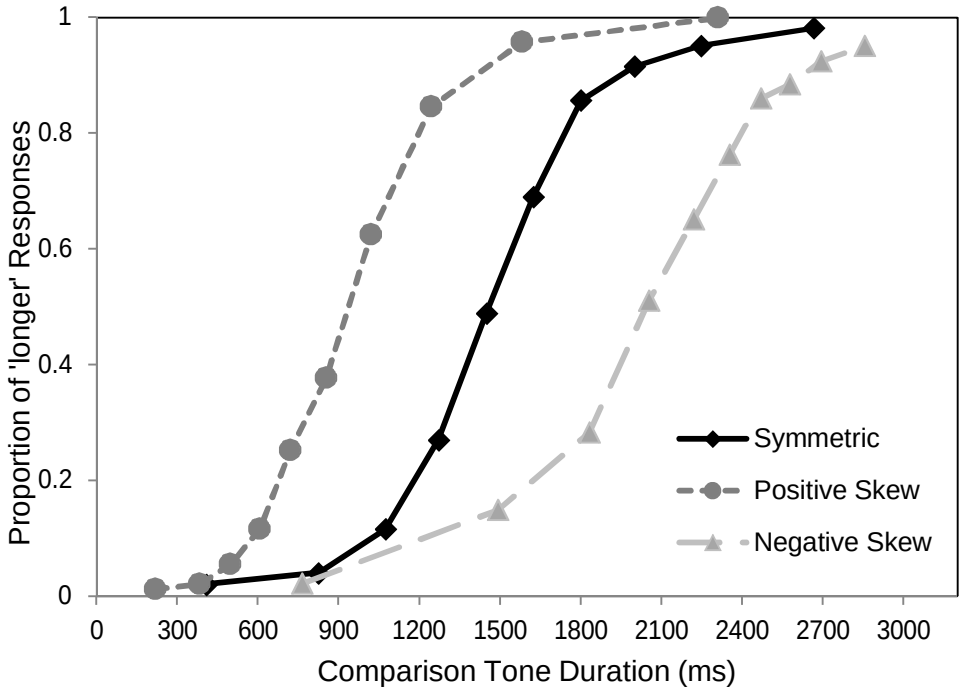

Figure 4. Psychophysical functions averaged over all participants, Experiment 3.

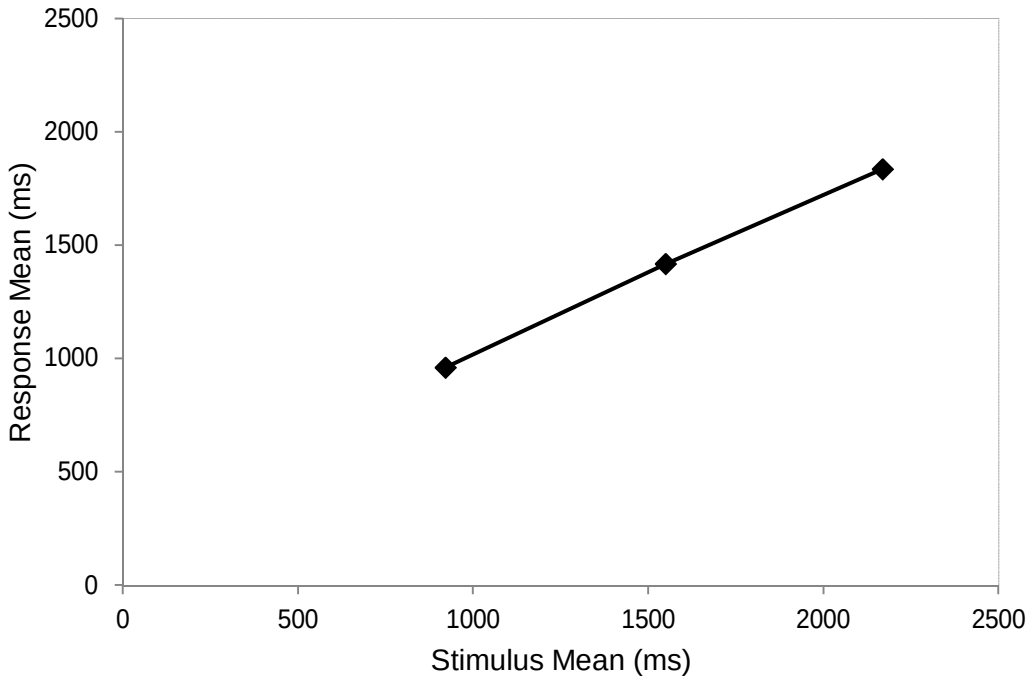

Figure 5. Response mean as a function of stimulus mean

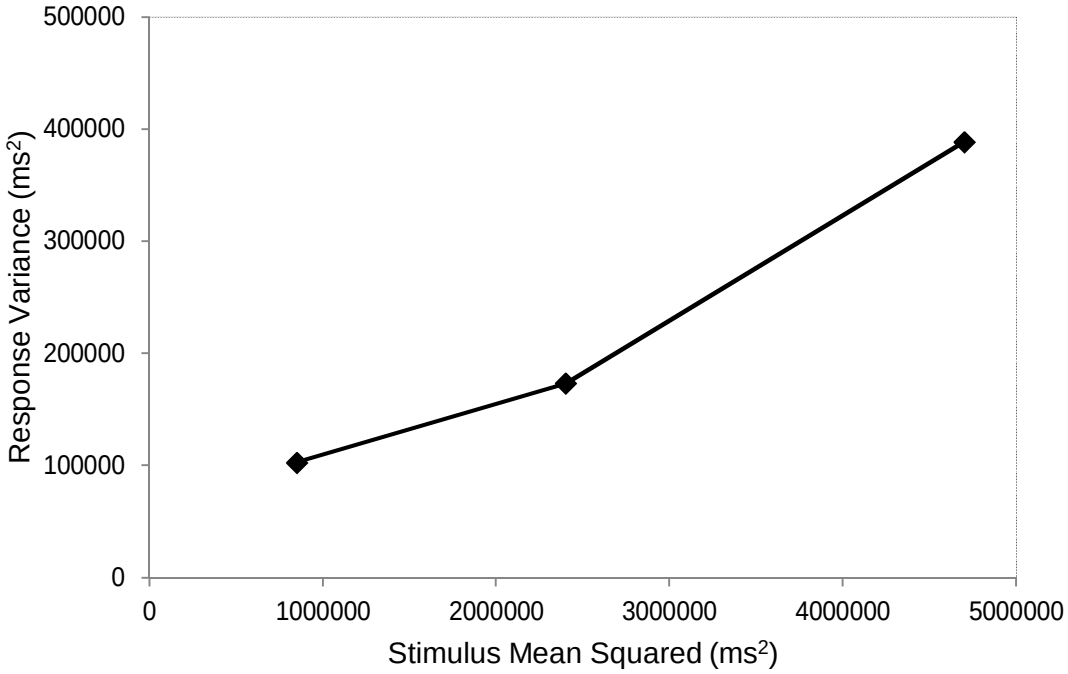

Figure 6. Response variance as a function of stimulus duration mean squared.