Title
The Interplay of Interval Models and Entrainment Models in Duration Perception

Author
Cheng, Tzu-Han

Publication Date
2020-07-02

Data Availability
The data associated with this publication are available at: https://osf.io/35H8N/

The Interplay of Interval Models and Entrainment Models in Duration Perception

Tzu-Han Zoe Cheng[1] and Sarah C. Creel[2]

[1] Department of Cognitive Science, UC San Diego, La Jolla, USA, tzcheng@ucsd.edu

[2] Department of Cognitive Science, UC San Diego, La Jolla, USA, sarah.c.creel@gmail.com

Abstract

Despite extensive research demonstrating the effect of temporal context on time perception, its underlying mechanisms remain poorly understood. One influential proposal to explain the temporal context effect is McAuley and Jones' (2003) framework that incorporates two classic timing models, interval and entrainment models. They demonstrated that listeners' duration estimates were shifted from reality in opposite directions when to-be-judged durations occurred earlier vs. later than an expected beat, which is predicted by their entrainment models. However, it is unclear about how long the entrainment lasts after the cessation of external stimulation. Here, we investigated the persistence of the entrainment effect in two experiments. In Experiment 1, we found that entrainment models predict the behaviors better after short delays (2 beats), while interval models predict better after long delays (4 beats). In Experiment 2, we extended the finding to a faster tempo and added one more delay length. Again, we found that entrainment was strongest after short delays (2 beats), while disappeared after medium (4 beats) and long delays (8 beats). Our findings suggest an interplay between entrainment and interval timings as a function of delays between successive events.

Keywords: time perception; entrainment; interval model; model comparison

PUBLIC SIGNIFICANCE STATEMENT: "This was the first study to test how long listeners expect sounds to occur "on the beat" after the end of a sequence of beats. Using a duration-judgment test, we found that beat sensitivity is present for about 2 cycles, but is mostly gone by 4 cycles after the end of a sequence of beats. More importantly, we found that duration judgments are still accurate even when the beat is gone, which suggests two different

processes may help listeners with making time judgments—one based on beats, the other based on long-term duration knowledge.”

Introduction

Time is a fundamental dimension of human perception, cognition, and action (see Grondin, 2010 for an overview). The scale of millisecond timing in particular has been intensely investigated to better understand human perceptual and motor capacity to process fine-grained temporal information in domains such as speech and music. The abilities both to estimate short time durations and to predict onset timings are regarded as important features of time perception ability (Coull, Cheng, & Meck, 2011). Estimating durations and predicting event onsets are not only essential for survival but also important in everyday life, allowing perceivers to safely cross busy intersections, synchronize with other dancers, play music with others, and understand others' speech by correctly perceiving temporal speech cues (e.g. voice onset time). The main goal of the current study is to understand the underlying mechanisms of human time perception in the scale of millisecond timing.

Although time perception has been studied for over a century, researchers are still debating its underlying mechanisms. Two prevalent classes of models, interval models and entrainment models, have been proposed to explain the wide range of time perception behaviors in short-time interval of millisecond timings. To avoid confusion of terminology, we would like to explicitly clarify here that we adopt the definitions of “entrainment models” and “interval models” used in McAuley and Jones (2003), especially in that we define interval models in our manuscript to be similar to clock-accumulator models, rather than the linear phase and period correction models used in Repp (2002). In research based on information processing theory, *interval models* introduce a stopwatch-like internal clock (Church, Meck, & Gibbon, 1994; Gibbon, 1977; Gibbon, Church, & Meck, 1984; Rakitin et al., 1998). One of the most representative interval models is the scalar expectancy theory (SET) (Allan, 1998; Gibbon 1971). In interval models, a pacemaker sends pulses through an attention-dependent switch to an accumulator. The accumulator counts the number of pulses in a certain duration, stores the duration code in working memory, then sends it to the long-term reference memory. Duration judgments are accomplished by comparing the new duration code in the working memory to the established duration code in the reference memory. It is important to note that interval models assume a full reset of the clock with each new stimulus onset.

On the other hand, *entrainment models* provide a different explanation by proposing an internal self-sustaining entrainable oscillator whose activation level presumably peaks at the onset of each external stimulus (Large & Jones, 1999; McAuley & Kidd, 1998). Large and Jones (1999) hypothesized peaks of oscillator activation to be peaks in attention. Their framework is borne out in findings of better auditory response accuracy for on-the-beat stimuli, recent findings of neural synchrony with exogenous rhythms (see Large, Herrera and Velasco 2015 for a review), as well as findings of superior *visual* discrimination at peaks (vs. troughs) of alpha oscillations (e.g. Mathewson et al., 2009, 2010; see also Calderone et al., 2014). With regard to context effects, entrainment models suggest that a change of the external rhythm will cause a disparity between the onset of external stimuli and the peak of the oscillator, which is evaluated by asking individuals to make duration judgments. If a stimulus (e.g. a pure tone onset, described by a black dot in Figure 1) in the driving rhythm occurs earlier than the driven oscillation, listeners will perceive the stimulus onset time as early, the tempo as faster, or the duration between the early stimulus and the next stimulus as shorter than it actually is. If an event in the driving rhythm occurs later than the driven oscillation, listeners will perceive the stimulus onset time as late, the tempo as slower, or the duration as longer. The key of entrainment models is the gradual correction/adjustment of the pace of the oscillator to match with the driving rhythm. Dynamic attending theory (DAT) is a generalization of entrainment theory, whereby the internal driven rhythm is conceptualized as an attentional pulse (Jones, 1976; Large & Jones, 1999). The gradual synchronization between internal driven rhythm and external stimuli onsets will narrow the attentional pulses. Eventually, the oscillator can precisely follow the external rhythm. Large and Snyder (2009) further extended the DAT model to neural resonance theory, hypothesizing that musical pulse and meter elicit corresponding neural rhythms which synchronize with external acoustic stimuli. This theory has been widely used in neural tagging research to capture how brains respond to acoustic rhythms (Nozaradan, Peretz, Missal, & Mouraux, 2011; Li et al., 2019).

Researchers have proposed different methods to reconcile these two models. One of the most influential proposals is McAuley and Jones' framework (2003). They applied the McAuley and Kidd (1998) paradigm (M&K paradigm), modified from the classic duration discrimination task, to test listeners' duration discrimination performances when the onset of the driving rhythm changes. Briefly, in a typical M&K task, participants first listen to a series of isochronous context tones to entrain them with a regular rhythm. Then they are asked to judge the relative length of two intervals: the standard interval (i.e. the time between a first pair of tones) and the comparison interval (i.e. the time between a second pair of tones). The onset time of the comparison pair is presented in expected (on-time), unexpected early, or unexpected late timing. Participants are instructed to indicate whether the second interval was shorter or longer than the first interval. Although both models predict effects of the context on duration judgments (please see Large & Jones, 1999 for the entrainment models; Drake and Botte, 1993 for the interval models), they have different predictions about the jitter of comparison onset time. Interval models predict no effect of the jitter of the onset of the comparison interval, that is, it does not matter if the comparison interval is aligned to the metrical grid defined by the context intervals. On the other hand, entrainment models predict an effect depending on this jitter because of the assumption that the driven rhythm will gradually adjust according to the driving rhythm. These models suggest that comparison intervals in the unexpected early condition should be judged as short, and comparison intervals in the unexpected late condition should be judged as long based on the calculation of the temporal contrast (illustrated in Figure 1). McAuley and Jones (2003) showed that listeners tended to respond "shorter" in the unexpected early condition and "longer" in the unexpected late condition, which is predicted by entrainment models, but not an interval model.

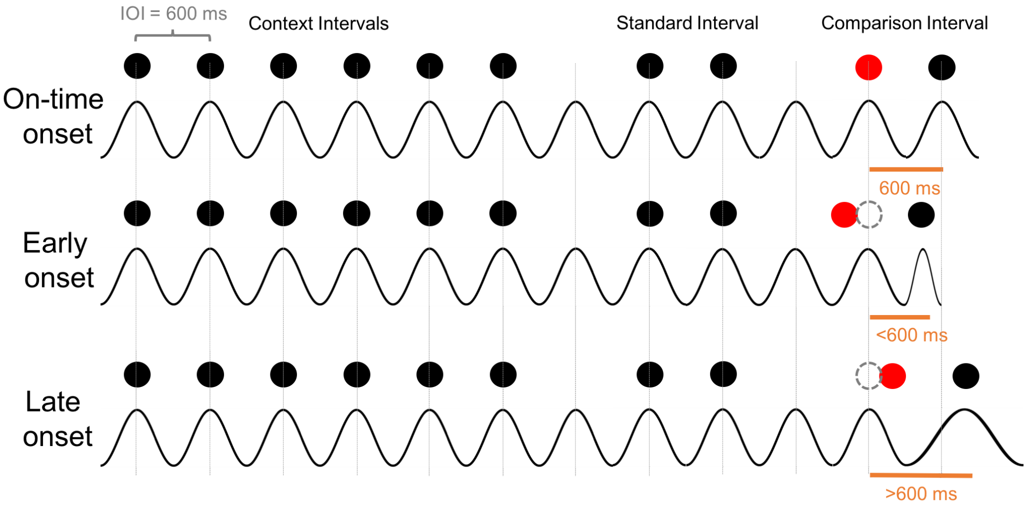

Figure 1. The schematic illustration of the durational judgment when the comparison interval is presented on-time, early and late. The solid black circles represent the explicit tones heard by the listeners. The solid red circles mark the first tone of the comparison interval, which is presented on-time (top row), early (middle) or late (bottom). The dashed circles in early and late onset represent the expected beat position. The waves represent the inner rhythm, or expectation, entrained by the explicit tones. The thick horizontal orange line indicates the subjectively perceived duration. Please note that the phase correction of the internal rhythm happens on the next cycle after the target tones.

McAuley and Jones (2003) developed a formula (please see the Model prediction section in the Methods and Materials and Appendix for more details) to simulate the prediction of two models by introducing linear phase-correction and period-correction terms to calculate the final temporal contrast. The phase is the disparity between the peak of the inner oscillator and the onset of a beat of the driving rhythm. Period represents the frequency, namely, the interval between two pulses of an inner oscillator. The weights of two parameters, phase and period, can be adjusted to use the same formula to simulate both interval models and entrainment models. Conceptually, interval models fully reset phase for each new stimulus in a series, while entrainment models enable more flexible dynamic changes of partial phase and partial period reset with a series of stimuli. Mathematically, this framework instantiates interval models by setting the phase parameter to one, marking the characteristic of full phase reset. In other words, interval models can be viewed as an extreme case of the entrainment models. McAuley and Jones (2003) favored the entrainment account in short timing intervals, however, their model did not specify whether and how long the entrainment persists after the cessation of external stimulation. Other features of auditory stimuli such as pitch height seem to decay over time. For example, Cowan, Saults, and Nugent (1997) first investigated and defined memory decay in a two-tone comparison task. They found that participants' performance was less accurate with an increasing delay length from 1.5 s to 12.0 s in between the two tones. Further, Snyder and colleagues (2008; Please see Snyder and Weintraub, 2013 for more details about the underlying mechanisms of declines in auditory memory) found that previous-context effects on auditory stream segregation based on pitch were diminished when the ISI changed from 1.44 s to 5.76 s. If entrainment is also based on similar sensory information, it is reasonable to expect that the entrainment effect would decay across time due to memory loss. So far, no one has investigated how long or even whether the entrainment effect persists over a longer delay, making it unclear how pervasive entrainment effects are. More importantly, though it has been argued that entrainment can sharpen perception (see Lakatos, Gross, & Thut, 2019 for a review), no study has investigated if duration judgement performance gets worse once the entrainment goes away or if it maintains at the same level because other underlying mechanisms compensate for the entrainment effect.

The current study

The current study asked whether and how long the entrainment effect persists over delay lengths. In Experiment 1, we kept M & K's context interval (600 ms) but modified the delay length between the standards and comparisons. This manipulation enables us to observe the entrainment effects in varied delay durations. The short delay (1200 ms, 2 beats based on the context interval) is equal to the delay length used in McAuley and Jones' Experiment 4. The long delay (2400 ms, 4 beats based on the context interval) is the double of the short delay. Following M&K, the participants were instructed to judge if the comparison interval is shorter or longer than the standard interval.

The primary goal of this experiment was to investigate which model works best to predict participants' performances in the short-delay vs. the long-delay condition. Note that we expect to replicate McAuley and Jones (2003) in our short-delay condition. If the entrainment effect persists for more than a few beats, as suggested by McAuley and Jones' framework, we should observe the same response pattern of responding shorter or longer in both the short-delay condition and the long-delay condition. If the entrainment effect decreases after the long delay, we will not observe the entrainment pattern in the long-delay condition. In this case, the performance accuracy would depend on whether the participants have access to the durational information. If the entrainment effect decays but a long-term reference memory of duration is robust, as suggested by interval models, then the accuracy should remain relatively high in the long-delay condition than the short-delay condition. On the other hand, if the entrainment decays and a long-term reference memory of duration is weak or absent, then the accuracy should be relatively low in the long-delay condition compared to the short-delay condition.

In Experiment 2, we aimed to replicate the effect of Experiment 1 and to extend our findings to a different absolute duration. We doubled the tempo (IOI=300 ms) of the context in Experiment 1 to observe the entrainment effects in varied tempi. Moreover, comparing the results of the Experiment 2 to Experiment 1 allowed us to test whether entrainment decreases after a certain amount of time (absolute delay) vs. a certain number of beats (relative delay).

Experiment 1

Materials and Methods

Ethics statement.

The IRB office of the University of California, San Diego (UCSD) approved this study. Participants signed informed consent forms to participate. The participants were given course credit for participating in the experiment.

Participants.

Our target sample size of 48 participants was determined as follows. We planned to look at three dependent measures: point of subjective equality (PSE), proportion short, and accuracy. If there is a change in degree of entrainment with the delay length, there should be an interaction of comparison onset time and delay length. In the absence of any prior test of delay length or its interaction, we used a conventional effect size of $\eta_p^2 = .10$, G*Power software (Faul, Erdfelder, Lang, & Buchner, 2007), which suggests a sample size of 46 to achieve power ≥ .80. We rounded this up to 48 to allow perfect counterbalancing of delay order and response key assignment.

To obtain our final sample size $N = 48$, we tested sixty-three participants (45 females, mean = 20.9 years old) from UCSD. Of these 63 participants, 15 were excluded for the following reasons:

one of them withdrew after the practice session, one of them had an age more than three standard deviations from the mean (Carrasco, Bernal, & Redolat, 2001; Coelho et al., 2004), three of them withdrew in partway through the maintask, and ten of them achieved less than 60% accuracy on the most distinguishable comparison intervals (the two comparison intervals which have the largest discrepancy from the standard intervals) in test trials, in one or both trial blocks. We describe our rationale for these exclusions below in the PSE in Results section; briefly, the replacement data did not change the significance patterns of accuracy or proportion short, but it substantially cleaned up the PSE effects. See Experiment 2 for use of the same criterion. After the exclusion, we analyzed the data from forty-eight participants (35 females, mean = 20.4 years old).

Stimuli and Procedure.

All stimulus sequences were comprised of a series of 60-ms 440-Hz sine tones, with various inter-tone intervals as described below. We let participants adjust to the most comfortable volume before they started the task. The stimuli were generated and delivered in MATLAB (MathWorks) using the Psychophysics Toolbox software (Brainard, 1997; Pelli, 1997).

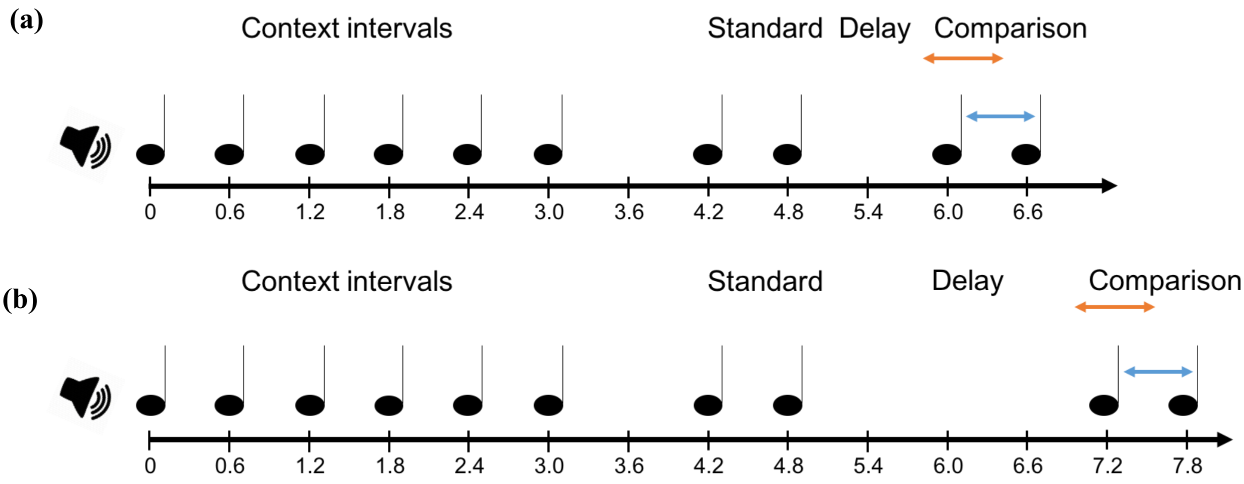

Figure 2. Schematic illustration of the duration discrimination paradigm in Experiment 1: (a) The short-delay condition. (b) The long-delay condition. The participants first heard six context tones with an inter-onset interval (IOI) of 600 ms. After a short silent period of 1200 ms, a pair of tones as standard interval and the other pair of tones as comparison interval were presented. The comparison

intervals could be 550, 570, 590, 610, 630, 650 ms (illustrated by the blue arrow). The onset time of the first tone of the comparison pair was manipulated to be 180 ms early, on-time or 180 ms late (illustrated by the orange arrow).

In the current experiment, we used a modified M & K paradigm (McAuley & Kidd, 1998), as McAuley and Jones (2003) used in their Experiment 4. On each trial (Figure 2), participants first heard a series of six context tones with an inter-onset interval (IOI) of 600 ms. After a short silent period of 1200 ms, a pair of tones as standard interval and the other pair of tones as comparison interval were presented. The standard interval was fixed at 600 ms. The comparison interval durations (550, 570, 590, 610, 630, and 650 ms) ranged from 50 ms shorter than the standard interval, increasing by 20 ms increments up to 50 ms longer than the standard interval. After each sequence (context + standard + comparison), the participant judged whether the comparison interval was shorter than or longer than the standard interval. As the main parameter of interest of this study, we manipulated the delay duration between the standard interval and the comparison interval, using 1200 ms +/- 180 ms for the short-delay and 2400 ms +/- 180 ms for the long delay. That is, for the short-delay, the first tone of the comparison interval would be presented on-time (1200 ms after the second tone of the standard interval), early (1020 ms after), or late (1380 ms after). For the *long delay*, the first tone of the comparison interval would be presented on-time (2400 ms after the second tone of the standard interval), early (2220 ms after), or late (2580 ms after). Please note that the shift of the comparison onset (+/- 180 ms) is used in McAuley and Jones (2003). This amount is large enough to be readily perceived and does not shift the comparison onset to any metrical position of the beats (i.e. 600 ms) or subdivision of the beat (i.e. 300 ms).

The participants were instructed to ignore the initial series of tones (i.e. the context intervals) preceding the standard interval and only focus on the standard and comparison intervals for answering each trial more accurately. The participants were asked to judge whether the duration of the comparison interval was shorter than or longer than the standard interval. They were instructed to press one of two keys on the keyboard to respond if the comparison interval was shorter or longer than the standard interval. The response time was self-paced.

A 2 (Delay Length) × 3 (Comparison Onset) × 6 (Comparison Duration) repeated measures design was used in this study. Following McAuley and Jones (2003), we leave out the comparison duration in our analyses because it is not relevant to our research question. To carefully control the response bias of responding by their dominant hand, we counterbalanced the response keys across participants (half of participants responded F key as shorter, J key as longer; the other half, J key as shorter, F key as longer). Short-delay and long-delay trials were blocked.

In this experiment, participants went through three blocks of trials: a brief practice block, and then two longer blocks with real trials. We fully counterbalanced the presentation order of the short-delay and long-delay blocks. In the practice block, twelve trials with short or long delay lengths and the two most-distinguishable comparison intervals (550 or 650 ms) were presented in random order. In the practice, we only presented trials with on-time comparison onset times. The practice block enabled participants to become familiar with the experimental task. Feedback indicating a correct or incorrect response appeared on the screen right after each response. During the practice block, participants were encouraged to adjust their performance based on the feedback. They were given up to 6 runs of the practice block (72 trials total) to achieve more than 75% accuracy in a run. They moved on to the real blocks once they had achieved higher than 75% accuracy in the practice block. All participants passed the practice accuracy criterion and continued to the main experiment, except for one who withdrew after the practice block.

In the second and third block, participants were presented with either the long delay duration or short delay duration with the three levels of comparison onset time and six levels of comparison intervals. The long-delay condition and short-delay condition were blocked and counterbalanced across participants. They could choose to take a break after every 54 trials. They had three breaks in the second block and three breaks in the third block. The only difference between the second and the third block was the manipulation of the delay length. The block with the short delay had 216 trials in total and took about 35 minutes to finish. The block with the long delay had 216 trials in total as well and took about 50 minutes to finish. In between the second and the third block, participants were instructed to complete two questionnaires about their basic demographic information and music background. Total years of music training did not interact with the other two variables of interest (i.e. Comparison Onset and Delay Length), suggesting that data patterns are not dependent on a high level of musical skill. The total experiment time was about 100 minutes, depending on the speed of the participant's responses and the length of the breaks.

Points of subjective equality (PSE).

We followed Macmillan and Creelman (1991, pp. 219–220) to calculate the points of subjective equality (PSE) for each participant in each condition. The PSE is the time interval subjectively considered as equally long as the standard interval, as operationalized by being judged as shorter 50% of the time. We first transformed the proportions of responding "short" for each of the six comparison intervals to z-score in the y-axis. Since z (1.00) is positive infinity and z (0.00) is negative infinity, we converted the 100 percent probability and 0 percent probability to 1-(1/2N) and 1/2N with N as the number of trials in each condition. Then we fit a first order (linear) polynomial function on the six z-transformed points. The intercept of the polynomial function on the x-axis represented the estimated PSEs for each condition. If a participant is responding with unbiased accuracy, their PSE should be 600 ms. If they are biased to think that the comparison duration is longer than it actually is, then their PSE should be smaller than 600 ms (because it will require a shorter comparison to sound equivalent to a standard duration of 600 ms). If they are biased to think the comparison is shorter than it actually is, then their PSE should be greater than 600 ms (because it will require a longer comparison to sound equivalent to a standard duration of 600 ms).

Model predictions.

In a M&K paradigm, interval models predict no effect of the jitter of the onset of the comparison interval. On the other hand, entrainment models predict that comparison intervals in the unexpected early condition should be judged as short, and comparison intervals in the unexpected late condition should be judged as long. Based on the framework of McAuley and Jones (2003), a timekeeper (which can be regarded as the internal clock in interval models or the oscillator in entrainment models) conducts linear phase and period corrections by calculating the temporal contrast with specific weights of phase (W_ϕ) and period (W_P)--that is, how much the model's internal phase and period are updated based on the observed external phase and period. (See Appendix for the calculation details). The model predictions of the subject's response are summarized in Table 1, using McAuley and Jones' (2003) parameters, which set the entrainment model to $W_\phi = 0.8$, $W_P = 0.05$ and the interval model to $W_\phi = 1$, $W_P = 0$.

Table 1

Predicting Values of Proportion Short Response and PSE from Interval ($W_\phi = 1$, $W_P = 0$*) and Entrainment Models (*$W_\phi = 0.8$, $W_P = 0.05$*). Note.* McAuley and Jones' (2003) formula predicts no difference between short and long-delay conditions in the proportion of responding that a comparison interval was shorter than the standard interval and PSE.

	Interval models		Entrainment models	
	Proportion short	PSE (ms)	Proportion short	PSE (ms)

Early	0.5	600	0.67	614.58
On-time	0.5	600	0.50	600.00
Late	0.5	600	0.33	585.42

Results

The effects we planned our study to detect are the effect of comparison onset time (early, on-time, late relative to an expected beat) and the interaction of onset time and delay length. The effect of comparison onset time is crucial for verifying that entrainment effects are present at short delays. The interaction term is crucial for determining if there is a decrease in entrainment at the longer delay. Each dependent measure--accuracy, proportion short, and point of subjective equality (PSE)--was analyzed in repeated-measures ANOVA. The independent variables were the three comparison onset times (early, on time, late) and the two delay lengths (short delay, long delay between standard and comparison). All significant interaction effects were investigated further via simple-effects comparisons. To break down 2-way interactions, we conducted post-hoc pairwise comparisons whose p-values were adjusted using the Bonferroni correction method (marked as p_b). In cases where there were sphericity violations, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. For thoroughness, we report relevant standard errors for each of these pairwise comparisons in Table 3 (see Franz & Loftus, 2012, on standard errors in within-subject designs).

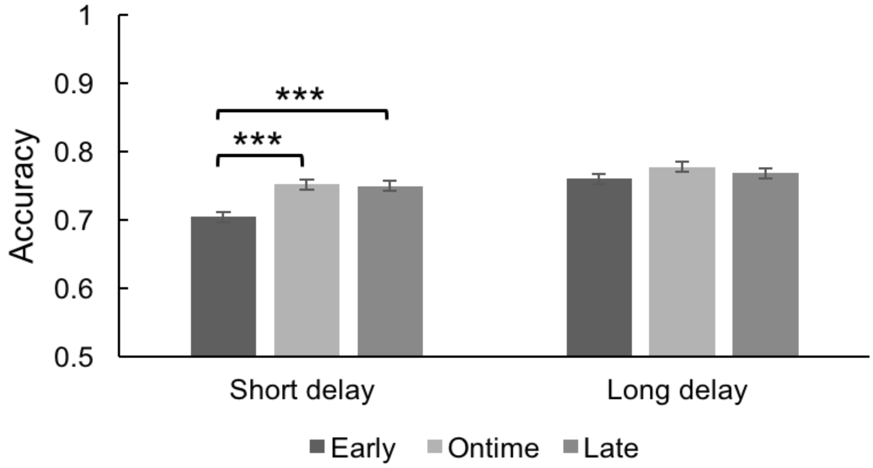

Figure 3. Proportion of correct response, collapsed across comparison interval duration. Error bars are standard errors, which reflect individual subject variability which is removed in within-subject ANOVAs and paired t-tests. *** $p_b < .0001$

Accuracy. We calculated the proportion of correct responses for all conditions (Figure 3). Consistent with a general effect of entrainment, there was a significant main effect of comparison onset time ($F(2, 94) = 14.11$, $p < .0001$, $\eta_p^2 = .23$), with lower accuracy for early comparison onsets compared to on-time comparison onsets ($t(47) = -4.65$, $p_b < .0001$, $d = -0.40$) and late comparison onsets ($t(47) = -4.83$, $p_b < .0001$, $d = -0.32$). There was a significant main effect of delay length ($F(1, 47) = 16.02$, $p < .0001$, $\eta_p^2 = .25$), with lower accuracy for short-delay trials. More interestingly, we found a significant interaction effect between comparison onset time and delay length ($F(2, 94) = 3.78$, $p < .05$, $\eta_p^2 = .07$). Simple effects analysis showed that in the short-delay condition, the comparison onset times significantly influenced the accuracy ($F(2, 94) = 19.21$, $p_b < .0001$, $\eta_p^2 = .29$). Specifically, in the short-delay condition, two out of three pairwise comparisons differed significantly, such that the accuracy of the early onset was lower than the on-time ($t(47) = -5.25$, $p_b < .0001$, $d = -0.54$) and late onsets ($t(47) = -5.05$, $p_b < .0001$, $d = -0.53$). There was no significant difference between on-time onset and late onset. By contrast, in the long-delay condition, comparison onset times did not affect accuracy ($F(2, 94) = 1.35$, $p_b = .53$, $\eta_p^2 = .03$).

Proportions of responding short. We calculated the proportion of responding short (Figure 4) to observe if there was a response bias to answer "shorter" in unexpected early onset trials and "longer" in unexpected late onset trials. There was a significant main effect of comparison onset time ($F(1.45, 68.06) = 83.02$, $p < .0001$, $\eta_p^2 = .64$), with more "short" responses for early comparison onsets compared to on-time ($t(47) = 10.37$, $p_b < .0001$, $d = 1.32$) and late comparison onset ($t(47) = 9.69$, $p_b < .0001$, $d = 1.54$), and more "short" responses for on-time comparison onsets compared to late comparison onset ($t(47) = -3.42$, $p_b < .005$, $d = 0.35$). This is consistent with a general entrainment effect. There was a significant main effect of delay length ($F(1, 47) = 5.36$, $p < .05$, $\eta_p^2 = .10$), with more "short" responses for short delays. More interestingly, we found a significant interaction effect between comparison onset time and delay length ($F(1.64, 77.10) = 62.73$, $p < .0001$, $\eta_p^2 = .57$). Simple effects analysis showed that in the short-delay condition, the comparison onset times significantly influenced the probability of responding "shorter" ($F(1.45, 68.05) = 98.98$, $p_b < .0001$, $\eta_p^2 = .68$), but in the long-delay condition there was no significant effect of comparison onset time. All pairwise comparisons in the short-delay condition differed significantly, such that the early onset was judged shorter more often than the on-time ($t(47) = 10.42$, $p_b < .0001$, $d = 1.82$) and late onsets ($t(47) = 10.82$, $p_b < .0001$, $d = 2.23$), and the on-time onset was judged shorter more often than the late onset ($t(47) = 4.53$, $p_b = .0001$, $d = 0.57$). By contrast, in the long-delay condition, comparison onset times did not affect the proportion of responding short ($F(1.76, 82.73) = 3.17$, $p_b = .11$, $\eta_p^2 = .06$). This revealed that participants tended to respond "shorter" in the early onset trials, while they responded "longer" in the late onset trials, regardless of the actual lengths of the comparison intervals. More importantly, this response tendency only happened in the short-delay condition: there were no significant pairwise comparisons in the long-delay condition.

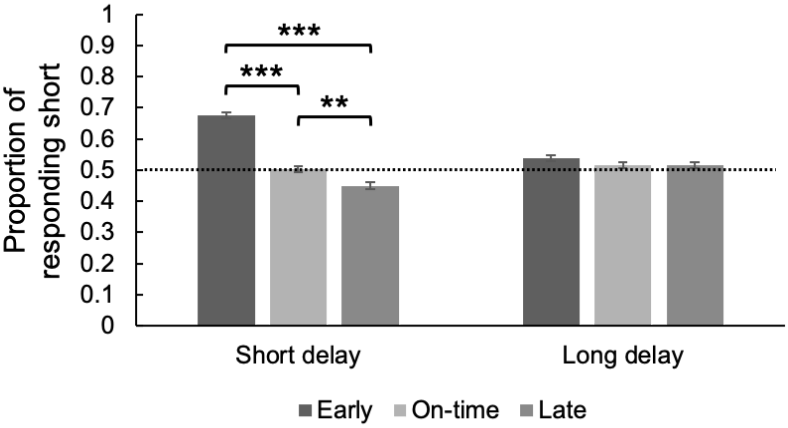

Figure 4. Proportion of responding that a comparison interval was shorter than the standard interval. The dashed line indicates the unbiased judgment (0.5) of the duration of comparison intervals. Error bars are standard errors, which reflect individual subject variability which is removed in within-subject ANOVAs and paired t-tests. ** $p_b < .001$, *** $p_b < .0001$

Points of subjective equality (PSE). To more clearly observe biases in perceived timing, we calculated the PSEs for each participant in each condition (Figure 5). If participants are perceiving veridically, then all PSEs should approximate 600 ms. However, if their duration perception is biased by comparison onset time, then PSEs should differ from 600 ms, as found by McAuley and Jones (2003) in their Experiment 4, which was equivalent to our short-delay condition. In our initial analyses of 48 participants, proportion short and accuracy data patterned as expected. However, we noticed that some participants were showing unusual PSE values, including two negative values (i.e. -1319 ms, -1864 ms) and multiple very long durations, such as 1839 ms, a value which seems implausible because it would mean that that participant would hear a comparison interval more than triple the duration of the preceding stimulus as being equivalent to the standard interval (600 ms). This was not an outcome that was reported in previous research, so we did not anticipate it. We traced this back to extremely low accuracy rates, and low accuracy generates very unstable PSE estimates. Since we were replicating McAuley and Jones (2003), who used PSE as their measure, it was crucial to get a clean PSE data set. Rather than excluding solely on the basis of what appeared to be aberrant PSEs, we opted to use what seemed like a reasonable accuracy criterion: whether participants have difficulty getting far above chance on the easiest (shortest and longest) stimuli. We chose 60% (exactly 10% above chance) as our cutoff value in both Experiment 1 and Experiment 2. We then replaced these low-accuracy participants, using this accuracy criterion for replacement participants as well. The average goodness-of-fit (R^2) of the linear model on converted PSE scores on the final sample was 0.84 (SD = 0.11). Table 2 compares major statistical outcomes of all three dependent measures between the original sample and the sample with replacement participants. As is clear from Table 2, this replacement only affected PSE.

Table 2

Comparison of all three dependent measures between original sample (N = 48) and the final sample (N = 48) with replacement participants.

Table 2a. Accuracy

	Main effect		Interaction effect
	Comparison onset	Delay length	Comparison × Delay
Original sample	$F(2, 94) = 7.43, p = .001, \eta_p^2 = .14$	$F(1, 47) = 5.81, p < .05, \eta_p^2 = .11$	$F(2, 94) = 3.51, p < .05, \eta_p^2 = .07$
Replacement sample	$F(2, 94) = 14.11, p < .0001, \eta_p^2 = .23$	$F(1, 47) = 16.02, p < .0001, \eta_p^2 = .25$	$F(2, 94) = 3.78, p < .05, \eta_p^2 = .07$

Table 2b. Proportion short

	Main effect		Interaction effect
	Comparison onset	Delay length	Comparison × Delay
Original sample	$F(1.32, 62.13) = 42.21, p < .0001, \eta_p^2 = .47$	$F(1, 47) = 6.68, p < .05, \eta_p^2 = .12$	$F(1.63, 76.66) = 45.77, p < .0001, \eta_p^2 = .49$
Replacement sample	$F(1.45, 68.06) = 83.02, p < .0001, \eta_p^2 = .64$	$F(1, 47) = 5.36, p < .05, \eta_p^2 = .10$	$F(1.64, 77.10) = 62.73, p < .0001, \eta_p^2 = .57$

Table 2c. PSE

	Main effect		Interaction effect
	Comparison onset	Delay length	Comparison × Delay
Original sample	$F(2, 94) = 1.45, p = 0.24, \eta_p^2 = .03$	$F(1, 47) = 1.93, p = .17, \eta_p^2 = .04$	$F(1.54, 72.33) = 2.11, p = .13, \eta_p^2 = .04$
Replacement sample	$F(1.18, 55.41) = 25.11, p < .0001, \eta_p^2 = .35$	$F(1, 47) = 7.93, p < .05, \eta_p^2 = .14$	$F(1.23, 57.68) = 33.71, p < .0001, \eta_p^2 = .42$

There was a significant main effect of comparison onset time ($F(1.18, 55.41) = 25.11, p < .0001, \eta_p^2 = .35$), with longer PSEs for early comparison onsets compared to on-time ($t(47) = 6.10, p_b < .0001, d = 1.12$) and late comparison onset ($t(47) = 4.91, p_b < .0001, d = 1.14$), consistent with an overall effect of entrainment. There was a significant main effect of delay length ($F(1, 47) = 7.93, p < .05, \eta_p^2 = .14$), with longer PSEs for short delays. These main effects were qualified by a significant interaction effect between comparison onset time and delay length ($F(1.23, 57.68) =$

33.71, $p < .0001$, $\eta_p^2 = .42$). Simple effects analysis showed that the comparison onset times significantly influenced the PSE in the short-delay condition ($F(1.16, 54.72) = 34.55$, $p_b < .0001$, $\eta_p^2 = .42$), but not in the long-delay condition ($F(1.34, 62.99) = 2.86$, $p_b = .17$, $\eta_p^2 = .06$). All pairwise comparisons were significant in the short-delay condition. The PSE of the early onset was longer than the on-time onset ($t(47) = 6.35$, $p_b < .0001$, $d = 1.21$) and late onset ($t(47) = 5.93$, $p_b < .0001$, $d = 1.40$), and the PSE in the on-time onset was longer than the late onset ($t(47) = 2.73$, $p_b < .05$, $d = 0.44$). Even though the simple effect of comparison onset time was not significant at the long delay, we analyzed pairwise differences nonetheless, to assess whether there were any remnants of entrainment. There was only one significant pairwise comparison, between early onset and on-time onset in the long-delay condition ($t(47) = 2.50$, $p_b < .05$, $d = 0.43$), such that the PSE of the early onset was longer than the on-time onset.

Exploratory analysis of PSE.

In order to determine whether timing biases persist at each delay, we compared subjective estimation (i.e. PSEs) to the standard interval (600 ms), as an “null point”, in each condition by a one-sample t test. This analysis confirmed that the response bias represented by PSEs was strongly evident in the short-delay condition: Comparison intervals that began on-time produced subjective estimates of the standard interval that did not differ statistically from the standard duration (mean PSE 600 ms, $t(47) = -0.06$, $p_b = 1$, $d = -0.01$), whereas participants tended to underestimate the comparison intervals that began unexpectedly early (mean PSE of 636 ms, $t(47) = 6.31$, $p_b < .0001$, $d = 0.91$) and overestimated those that began late (mean PSE of 592 ms, $t(47) = -2.73$, $p_b < .05$, $d = -0.39$). (Note that responding “shorter” more frequently corresponds to longer PSE and responding “longer” more frequently corresponds to shorter PSE due to the definition and the calculation of PSE.) On the other hand, this trend was weak in the long-delay condition (mean PSE of 607 ms, 601 ms, 601 ms), which showed a slight underestimate of early onset trials ($t(47) = 2.83$, $p_b < .05$, d

= 0.41), but estimation that was indistinguishable from correct for both on-time ($t(47) = 0.67$, p_b = 1) and late onset trials ($t(47) = 0.41$, p_b = 1).

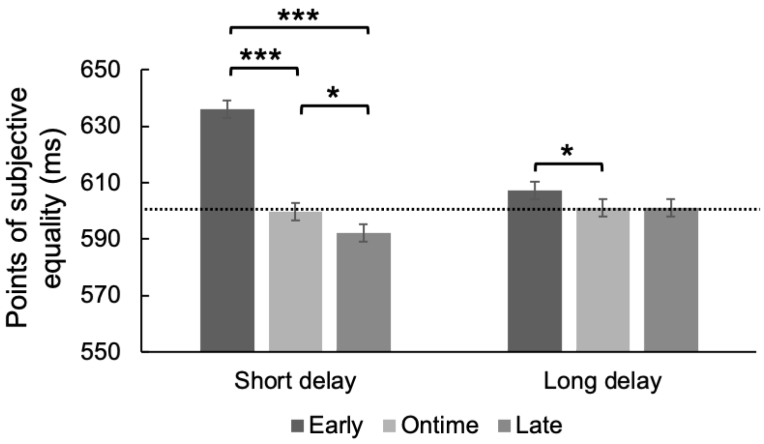

Figure 5. Points of subjective equality (PSE). The dashed line indicates the unbiased PSE (600ms). Error bars are standard errors, which reflect individual subject variability which is removed in within-subject ANOVAs and paired t-tests. * $p_b < .05$, *** $p_b < .0001$

Table 3

Mean difference scores and standard errors for within-subjects pairwise comparisons. Standard errors appear in parentheses.

	Short delay			Long delay		
	Early – Ontime	Early – Late	Ontime – Late	Early – Ontime	Early – Late	Ontime – Late
Accuracy	-0.05 (0.01)	-0.05 (0.01)	0.00 (0.01)	-0.02 (0.01)	-0.01 (0.01)	0.01 (0.01)
Proportion short	0.17 (0.02)	0.22 (0.02)	0.05 (0.01)	0.02 (0.01)	0.02 (0.01)	-0.00 (0.01)
PSE	36.12 (5.69)	43.93 (7.40)	7.80 (2.86)	6.29 (2.51)	6.07 (3.89)	-0.22 (2.30)

Discussion

In Experiment 1, we successfully replicated McAuley and Jones (2003)'s Experiment 4 in our short-delay condition. Like McAuley and Jones (2003), we observed results that were consistent with the prediction of entrainment models: Listeners tended to judge the comparison intervals as being shorter than the standard intervals when the onset time of the comparison interval was early, and longer than the standard when late. This response bias was reflected both in the proportion of responding short and PSE, with the highest proportion short (longest PSE) in the early trials, intermediate in the on-time trials, and lowest (shortest PSE) in the late trials (Figures 4 and 5).

Our second goal was to determine if entrainment models or interval models more precisely predicted listeners' performance in the long-delay condition. The winner appears to be interval models: we observed a near-disappearance of the entrainment effect in the long-delay condition, which has not yet been studied. Compared to the short-delay condition, listeners tended to judge the comparison intervals more accurately, without strong biases to report the comparison interval as shorter or longer. The proportion of responding short revealed a flat line close to 50 percent, and PSE revealed a nearly-flat line close to 600 ms across three onset times. Furthermore, although the entrainment effects only influenced the duration judgement task at short delays, we observed even higher accuracy in long-delay (76.8 percent correct) than short-delay (73.5 percent correct) conditions. This result indicates that entrainment might not be the only source of information that aids listeners in discriminating durations. With regard to our hypothesis, if the entrainment effect decreases across time and we still observe a similar or even higher level of accuracy, it is likely that the reference memory assumed in interval models plays a role in the duration discrimination. Another possible explanation of the slightly higher accuracy in the long-delay condition could be due to better attentional preparation allowed by the longer foreperiod as found by Grondin & Rammsayer (2003). In sum, we speculate that the high accuracy at both delays is that listeners allocate attention in the short timing interval as suggested in the Dynamic Attending Theory, but depend more on attentional preparation and/or an interval reference memory in the long-delay condition.

In sum, we found a duration judgement pattern similar to the prediction of the entrainment models when there was a short delay between standard and comparison intervals, and a pattern similar to the predictions of interval models when there was a long delay between standard and comparison intervals. Our data suggest that the entrainment effect may disappear somewhere between 2 to 4 beats (1200 to 2400 ms). To extend our findings to a different tempo and to further investigate whether entrainment decays after a certain delay time (absolute time match) or after a certain number of beats (relative time match), we ran Experiment 2. This experiment was pre-registered on Open Science Framework (https://osf.io/nky52) before data collection.

Experiment 2

Materials and Methods

Ethics statement. Identical to Experiment 1.

Participants.

Our estimated effect size, and thus our target sample size, was identical to Experiment 1. Participants were recruited in the manner described for Experiment 1. To obtain our final sample size $N = 48$, we tested 56 participants (38 females, mean = 20.2 years old). Of these 56 participants, 4 of them withdrew partway through the main task; 1 of them reversed the response keys during the experiment; 3 of them achieved less than 60% accuracy on the two most distinguishable comparison intervals in one or more blocks of test trials (see Materials and Methods in Experiment 1 and our preregistration for more details about this stop rule). After the exclusion, we analyzed the data from forty-eight participants (31 females, mean = 20.1 years old).

Stimuli and Procedure.

The stimuli, procedure and design matched Experiment 1 except for two changes. First, we doubled the tempo of the task, that is, set the IOI to 300 ms instead of 600 ms. Second, we used three delay lengths instead of two: 600 ms, 1200 ms, and 2400 ms, all ± 90 ms to account for the early and late onsets. The delay between context and standard was 600 ms. The standard interval was fixed at 300 ms. The comparison interval durations (275, 285, 295, 305, 315 and 325 ms) ranged from 25 ms shorter than the standard interval, increasing by 10 ms increments up to 25 ms longer than the standard interval (consistent with doubling the tempo; Experiment 1 comparisons ranged from -50 to +50 ms). As before, the participant judged whether the comparison interval was shorter than or longer than the standard interval (Figure 6). As the main interest in the current study, again, the short-delay, medium-delay and long-delay trials were blocked, and block order was fully counterbalanced across participants. Each block had the same number of trials as blocks in Experiment 1 (i.e. 216 trials) but took less time (~30 minutes) to finish due to the doubled tempo. Participants were asked to complete two questionnaires about their basic demographic information and music background after the task. We found that the music training years do not interact with the two variables of interest, suggesting that entrainment decay is not limited to musically-skilled individuals. The total experiment time was about 120 minutes, depending on the speed of their responses and the length of the breaks.

A 3 (Delay Length) × 3 (Comparison Onset) × 6 (Comparison Duration) repeated measures design was used in Experiment 2. We followed the same method to calculate PSE and all analyses as we did for Experiment 1.

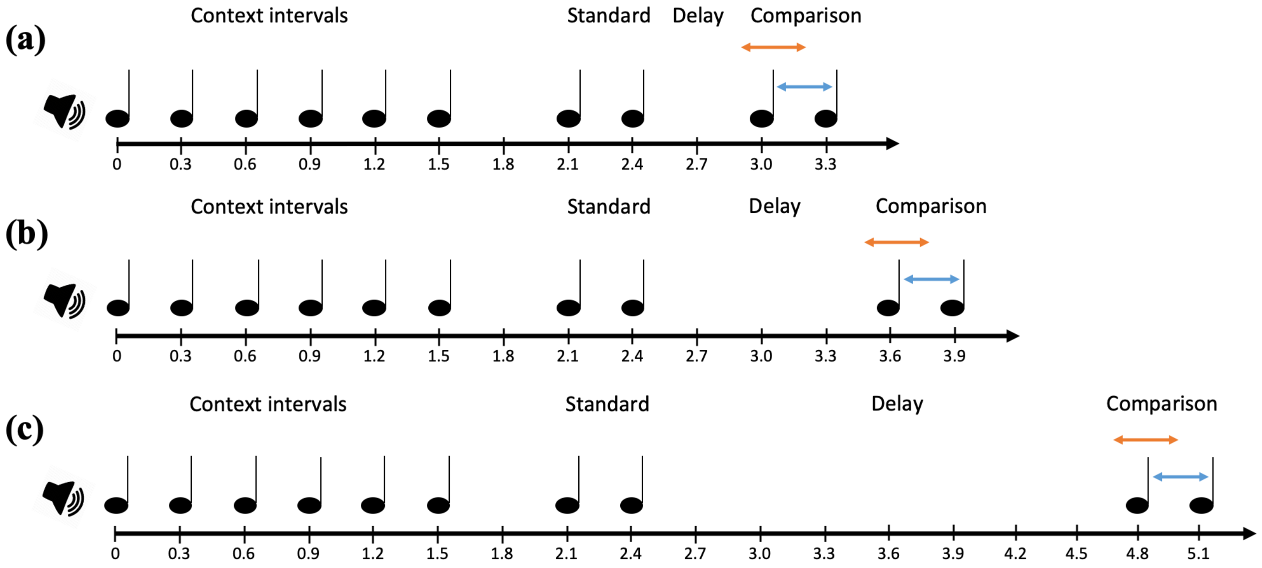

Figure 6. Schematic illustration of the duration discrimination paradigm in Experiment 2: (a) The short-delay condition. (b) The medium-delay condition. (c) The long-delay condition. The participants first heard six context tones with an inter-onset interval (IOI) of 300 ms. After a short silent period of 600 ms, a pair of tones as standard interval and the other pair of tones as comparison interval were presented. The comparison intervals could be 275, 285, 295, 305, 315, 325 ms (illustrated by the blue arrow). The onset time of the first tone of the comparison pair was manipulated to be 90 ms early, on-time or 90 ms late (illustrated by the orange arrow).

Results

Analyses mirrored those in Experiment 1.

Accuracy. We calculated the proportion of correct responses for all conditions (Figure 7). Consistent with an overall entrainment effect, there was a significant main effect of comparison onset time ($F(2, 94) = 30.31$, $p < .0001$, $\eta_p^2 = .39$), with lower accuracy for early comparison onsets compared to on-time comparison onsets ($t(47) = -7.53$, $p_b < .0001$, $d = -0.59$) and late comparison onsets ($t(47) = -6.08$, $p_b < .0001$, $d = -0.61$). There was a significant main effect of delay length ($F(2, 94) = 91.25$, $p < .0001$, $\eta_p^2 = .66$), with lower accuracy for short-delay trials compared to medium-delay trials ($t(47) = -9.56$, $p_b < .0001$, $d = -1.28$) and long-delay trials ($t(47) = -13.23$, $p_b <$

.0001, d = -1.70), and lower accuracy for medium-delay trials compared to long-delay trials ($t(47)$ = -2.76, $p_b < .05$, d = -0.33). Comparison onset time and delay length interacted ($F(4, 188) = 21.03$, $p < .0001$, $\eta_p^2 = .31$). Simple effects analysis showed that in the short-delay condition, the comparison onset times significantly influenced the accuracy ($F(2, 94) = 51.65$, $p_b < .0001$, $\eta_p^2 = .52$): accuracy of the early onset was lower than the on-time ($t(47) = -9.33$, $p_b < .0001$, d = -1.19) and late onsets ($t(47) = -8.45$, $p_b < .0001$, d = -1.36). There was no significant difference between on-time onset and late onset. By contrast, comparison onset times did not affect accuracy in the medium ($F(2, 94)$ = 2.58, $p_b = .24$, $\eta_p^2 = .05$) and long-delay conditions ($F(2, 94) = 0.29$, $p_b = 1$, $\eta_p^2 = .00$).

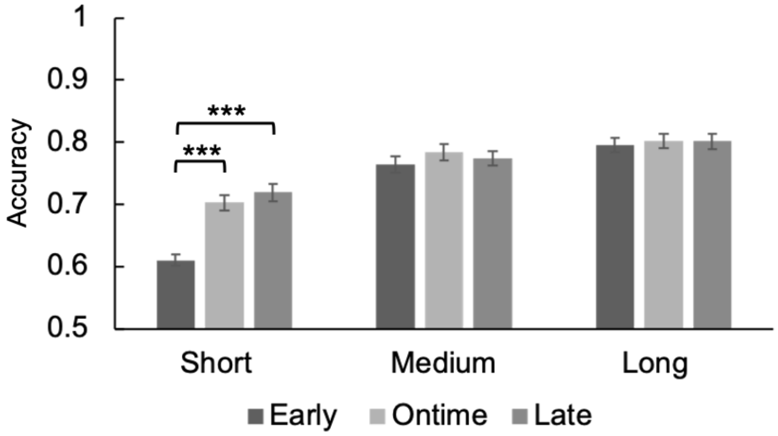

Figure 7. Proportion of correct response, collapsed across comparison interval duration. Error bars are standard errors, which reflect individual subject variability which is removed in within-subject ANOVAs and paired t-tests. *** $p_b < .0001$

Proportions of responding short. For proportion of responding short (Figure 8), there was a significant main effect of comparison onset time ($F(2, 94) = 227.60$, $p < .0001$, $\eta_p^2 = .83$), with more “short” responses for early comparison onsets compared to on-time ($t(47) = 15.47$, $p_b < .0001$, d = 1.36) and late comparison onset ($t(47) = 18.44$, $p_b < .0001$, d = 1.85), and more “short” responses for on-time comparison onsets compared to late comparison onset ($t(47) = 5.32$, $p_b <$ .0001, d = 0.38). This is consistent with an overall effect of entrainment. There was a significant main effect of delay length ($F(2, 94) = 12.86, p < .0001, \eta_p^2 = .21$), with more "short" responses for short-delay trials compared to medium-delay trials ($t(47) = 3.39, p_b < .01, d = 0.36$) and long-delay trials ($t(47) = 4.84, p_b < .0001, d = 0.52$). Finally, we found a significant interaction effect between comparison onset time and delay length ($F(2.73, 128.08) = 163.68, p < .0001, \eta_p^2 = .78$). Simple effects analysis showed that the comparison onset times significantly influenced the probability of responding "shorter" in the short-delay condition ($F(1.77, 83.08) = 257.56, p_b < .0001, \eta_p^2 = .85$) and the medium-delay condition ($F(2, 94) = 16.97, p_b < .0001, \eta_p^2 = .27$), but in the long-delay condition there was no significant effect of comparison onset time. All pairwise comparisons in the short-delay condition differed significantly, such that the early onset was judged shorter more often than the on-time ($t(47) = 16.02, p_b < .0001, d = 2.56$) and late onsets ($t(47) = 19.32, p_b < .0001, d = 3.59$), and the on-time onset was judged shorter more often than the late onset ($t(47) = 5.98, p_b < .0001, d = 0.71$). In the medium-delay condition, the early onset was judged shorter more often than the on-time ($t(47) = 3.38, p_b < .01, d = 0.32$) and late onsets ($t(47) = 6.00, p_b < .0001, d = 0.52$). By contrast, in the long-delay condition, comparison onset times did not affect the proportion of responding short ($F(2, 94) = 0.56, p_b = 1, \eta_p^2 = .01$).

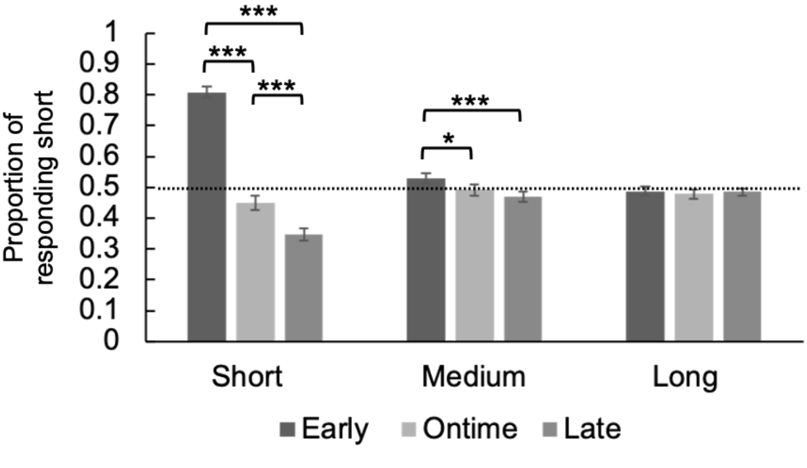

Figure 8. Proportion of responding that a comparison interval was shorter than the standard interval. The dashed line indicates the unbiased judgment (0.5) of the duration of comparison intervals. Error bars are standard errors, which reflect individual subject variability which is removed in within-subject ANOVAs and paired t-tests. * $p_b < .05$, *** $p_b < .0001$

Points of subjective equality (PSE). To more clearly observe biases in perceived timing, we calculated the PSEs for each participant in each condition (Figure 9). If participants are perceiving veridically, then all PSEs should approximate 300 ms. If their duration perception is biased by comparison onset time, then PSEs should differ from 300 ms. The average goodness-of-fit (R^2) of the linear model on converted PSE scores was 0.81 (SD = 0.09). One participant showed an unusual PSE value, -188000000000000000, in early trials of the short-delay condition. We excluded this subject for statistical analysis of PSE. Consistent with an overall effect of entrainment, comparison onset time was significant ($F(1.07, 49.29) = 36.5$, $p < .0001$, $\eta_p^2 = .44$), with longer PSEs for early comparison onsets compared to on-time ($t(46) = 5.99$, $p_b < .0001$, $d = 0.98$) and late comparison onset ($t(46) = 6.18$, $p_b < .0001$, $d = 1.19$), and longer PSEs for on-time comparison onsets compared to late comparison onset ($t(46) = 3.97$, $p_b < .001$, $d = 0.40$). There was a significant main effect of delay length ($F(1.07, 49.20) = 8.53$, $p < .01$, $\eta_p^2 = .16$), with longer PSEs for short-delay trials compared to medium-delay trials ($t(46) = 2.83$, $p_b < .05$, $d = 0.51$) and long-delay trials ($t(46) = 3.05$, $p_b < .05$, $d = 0.58$). These main effects were qualified by a significant interaction effect between comparison onset time and delay length ($F(1.08, 49.48) = 33.60$, $p < .0001$, $\eta_p^2 = .43$). Simple effects analysis showed that the comparison onset times significantly influenced the PSE in the short-delay condition ($F(1.05, 48.28) = 35.11$, $p_b < .0001$, $\eta_p^2 = .43$) and the medium-delay condition ($F(2, 92) = 9.31$, $p_b < .001$, $\eta_p^2 = .17$), but there was not a significant effect of the comparison onset times on the long-delay condition ($F(2, 92) = 1.17$, $p_b = .94$, $\eta_p^2 = .03$). All pairwise comparisons were significant in the short-delay condition: the PSE of the early onset was longer than the on-time onset ($t(46) = 5.96$, $p_b < .0001$, $d = 1.05$) and late onset ($t(46) = 5.98$, $p_b <$

.0001, $d = 1.23$), and the PSE in the on-time onset was longer than the late onset ($t(46) = 3.80$, $p_b < .01$, $d = 0.50$). In the medium-delay condition, early was longer than late ($t(46) = 4.37$, $p_b < .001$, $d = 0.50$). There were no significant pairwise differences of the comparison onset times in the long-delay condition.

Exploratory analysis of PSE.

In order to determine whether timing biases persist at each delay, we compared subjective estimation (i.e. PSEs) to the standard interval (300 ms), as a “null point”, in each condition by a one-sample t test. This analysis confirmed that the response bias represented by PSEs was strongly evident in the short-delay condition: Comparison intervals that began on-time produced subjective estimates of the standard interval that did not differ statistically from the standard duration (mean PSE 295 ms, $t(46) = -1.33$, $p_b = .57$, $d = -0.19$), whereas participants tended to underestimate the comparison intervals that began unexpectedly early (mean PSE of 363 ms, $t(46) = 4.90$, $p_b < .0001$, $d = 0.71$) and overestimated those that began late (mean PSE of 284 ms, $t(46) = -5.42$, $p_b < .0001$, $d = -0.79$). (Note that responding “shorter” more frequently corresponds to longer PSE and responding “longer” more frequently corresponds to shorter PSE due to the definition and the calculation of PSE.) On the other hand, this pattern disappeared in the medium (mean PSE of 303 ms, 301 ms, 298 ms) and long-delay conditions (mean PSE of 300 ms, 299 ms, 299 ms).

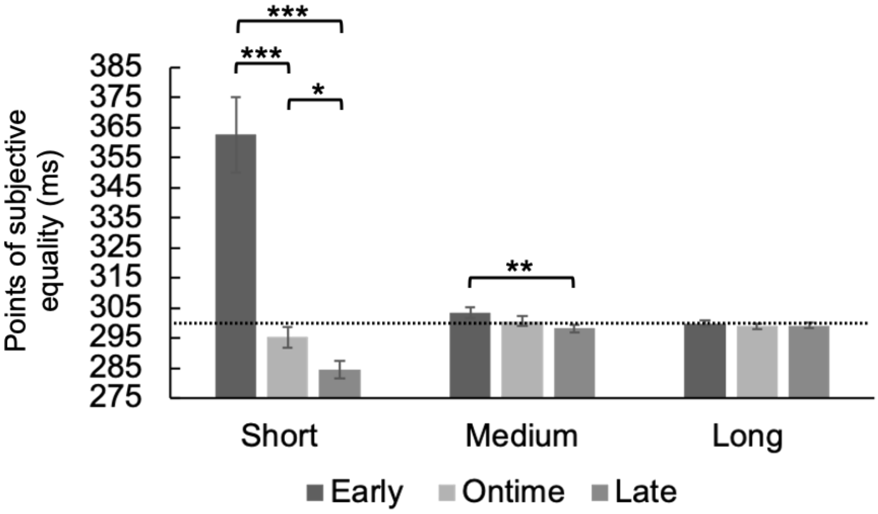

Figure 9. Points of subjective equality (PSE). The dashed line indicates the unbiased PSE (600ms). Error bars are standard errors, which reflect individual subject variability which is removed in within-subject ANOVAs and paired t-tests. * $p_b < .05$, ** $p_b < .001$, *** $p_b < .0001$

Table 4

Mean difference scores and standard errors for within-subjects pairwise comparisons. Standard errors appear in parentheses. Note that we excluded the subject who had an aberrant PSE value for the mean difference and standard error for PSE.

	Short delay			Medium delay			Long delay		
	Early – Ontime	Early – Late	Ontime – Late	Early – Ontime	Early – Late	Ontime – Late	Early – Ontime	Early – Late	Ontime – Late
Accuracy	-0.09 (0.01)	-0.11 (0.01)	-0.02 (0.01)	-0.02 (0.01)	-0.01 (0.01)	0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	0.00 (0.01)
Proportion short	0.36 (0.02)	0.46 (0.02)	0.10 (0.02)	0.04 (0.01)	0.06 (0.01)	0.02 (0.01)	0.01 (0.01)	0.00 (0.01)	-0.01 (0.01)
PSE	67.22	78.14	10.92	2.71	5.15	2.44	0.89	0.69	-0.20

	(11.15)	(12.93)	(2.84)	(1.24)	(1.17)	(1.13)	(0.51)	(0.68)	(0.60)

Discussion

Our first goal was to replicate the observed decay of entrainment in Experiment 1 in a different tempo (600 ms IOI vs 300 ms IOI). We achieved this goal: in Experiment 2, we again observed statistically significant entrainment effects in the short delay, but not in the longer delays. Similar to what we found in Experiment 1, we also observed a higher accuracy in long-delay than medium-delay condition, and a higher accuracy in medium-delay than short-delay condition. Again, this suggests that entrainment, rather than sharpening duration estimation, might actually deteriorate subjects' duration estimation, especially in the early trials in the short-delay condition, with some other source of information accurately guiding duration judgments at longer delays.

Our second goal was to test over what time scale entrainment decreases: after a certain amount of time (absolute delay), or a certain number of beats (relative delay)? Experiment 1 had delays of 1200 ms (2 beats) and 2400 ms (4 beats), while Experiment 2 had delays of 600 ms (2 beats), 1200 ms (4 beats) and 2400 ms (8 beats). If the absolute delay account is true, we should observe entrainment disappearing between 1200 ms and 2400 ms, as found in our previous experiment. If the relative delay account is true, we should observe entrainment disappearing between 600 ms (2 beats) and 1200 ms (4 beats). Here, we found that entrainment disappears somewhere between 600 ms (2 beats) and 1200 ms (4 beats), which suggests a relative account rather than an absolute account. However, the magnitude of the entrainment effect at 2 beats here appears to be much stronger than Experiment 1. This would seem to be more in favor of entrainment decay over absolute time, but it is not clear whether one can really equate across the two experiments, both in terms of effect magnitude and in terms of slightly different populations tested since studies were run at two different time points. We return to this point in the General

Discussion.

General Discussion

The current research aimed to assess the persistence of entrainment effects on duration perception in two experiments. Accordingly, we gauged how well interval models vs. entrainment models of time perception predict listeners' duration comparison performance with a range of interstimulus intervals (delay durations). The task required participants to compare the duration of a comparison interval to a standard interval. We manipulated the early, on-time and late comparison starts with varied delays between the standard and comparison intervals. In Experiment 1, our data suggest that entrainment models predict the behaviors better after short delays (2 beats, 1200 ms), while interval models predict the behaviors better after long delays (4 beats, 2400 ms). In Experiment 2, we extended the finding with doubled presentation rate (Experiment 1: 600 ms IOI vs. Experiment 2: 300 ms IOI) and one more delay length (Experiment 1: short, long delay vs. Experiment 2: short, medium, long delay). In Experiment 2, we found that entrainment models predict the behaviors after short delays, while interval models predict the behaviors better after medium and long delays.

A different question concerns the time scale over which entrainment disappears: does it vanish over an amount of time or a certain number of beats? Based on patterns of statistical significance in both PSE and proportion short, the short-delay conditions in Experiment 2 matched the short-delay condition in Experiment 1 in showing entrainment effects. The medium and long-delay conditions in Experiment 2 matched the long-delay condition in Experiment 1, suggesting lack of entrainment but an interval timing effect. However, the amount of response bias, most clearly visualized in the proportion of responding short, was quite different between the two experiments (Figure 10). To further validate the relative over the absolute decay account, it is critical to compare the effect size of the medium-delay condition in Experiment 2 (1200 ms) with the short-delay condition in Experiment 1 (also 1200 ms) (EXP2 medium vs EXP1 short) because the delay length was controlled to be the same, the only difference was set by the *beats* during the delay (4 beats vs 2 beats). If entrainment decay happens in absolute time, then the effect of response bias after the same delay length should be identical. This is not what we found; instead, we found a smaller effect size in EXP2 medium ($\eta_p^2 = .27$) than the EXP1 short ($\eta_p^2 = .68$). This finding demonstrated that entrainment decays in a beat-based manner, rather than a time-based manner. However, please note that the effect size of the EXP 2 short ($\eta_p^2 = .85$) was stronger than the EXP 1 short ($\eta_p^2 = .68$), suggesting an effect of absolute time such that entrainment is stronger after the short absolute delay. This suggests that although entrainment disappears after a certain number of beats, the decay rate is influenced by more complex processing that might include the nature of memory loss over time according to different tempo, and perhaps the entire length, of the entraining context. Sorting this out is beyond the scope of the current study, but future research could further explore the joint influences of tempo, total context duration, number of context events, perceived meter, and other parameters on entrainment decay, as well as further investigating entrainment decay as a function of memory loss.

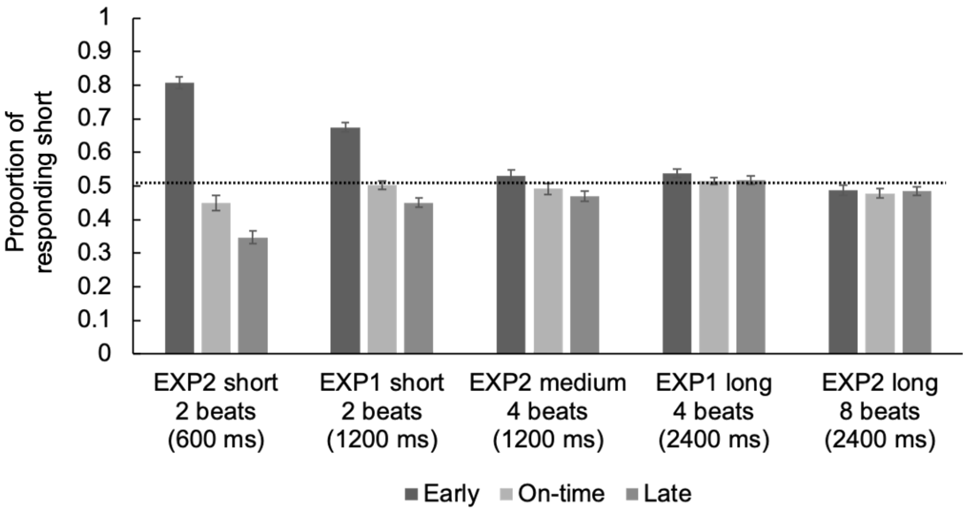

Figure 10. Proportion of responding that a comparison interval was shorter than the standard interval in Experiment 1 and Experiment 2. The dashed line indicates the unbiased judgment (0.5) of the duration of comparison intervals. Error bars are standard errors.

An interesting pattern observed in our results for both experiments is that our participants showed a stronger response bias in early onset trials than the late onset trials in both delay lengths, as did McAuley and Jones (2003) in their Experiment 4. The asymmetry of the response bias cannot be explained or predicted by either the interval or the entrainment models laid out by McAuley and Jones (2003). Nevertheless, this finding is consistent with Repp's (1998a) finding that late trials have higher accuracy than early trials, which suggests that listeners more easily detect the duration changes with a late onset than an early onset (see also Repp, 2002, who found that phase correction responses are larger when a stimulus occurred unexpectedly early than when it was unexpectedly late). Repp (1998b) suggested that the reason for this is that a stimulus which occurs earlier than expected will precipitate the oscillator to start phase and period correction faster than a stimulus which occurs later than expected. Similar effects have been found in actual musical practice. For instance, there is often a slowing down of the tempo at the end of a music phrase or the end of a whole music piece, which is a frequently applied performance practice called *ritardando* (Repp, 1992; Honing, 2003). This has been linked to a phenomenon observed widely across languages called *prepausal lengthening*, in which there is an increase in speech duration just prior to the end of a phrase (Umeda, 1975; Klatt, 1976; Crystal & House, 1988; Campbell & Isard, 1991).

Another interesting explanation, which our study did not test directly, is based on the time-order error (TOE) effect. This is a systematic error in which the comparison duration is *underestimated* (i.e. judged shorter than it actually is) after shorter delays, but this underestimation decreases with a longer delay between standard and comparison (Allan, 1979). One could argue that our data fit this pattern; however, if TOE decreases with delay length, one would predict a *monotonic* decrease between the shortest delay (early trials in short-delay condition) and the longest delay (late trials in long-delay condition). This is not consistent with what we found, which is more like a damped oscillation (see Figures 4, 5, 8, 9). In Experiment 1, the late trials in the short-delay condition were *overestimated* more than the early trials in long-delay condition, even though it has a shorter delay length. A similar pattern was found in Experiment 2. These observations suggest that on top of the entrainment timing and interval timing system, the well-known TOE might contribute to the stronger response bias in early onset trials but cannot explain the full data pattern.

Limitation.

In sum, we concluded that participants' time perception was better predicted by entrainment models if the delay length is short, while their performances were better predicted by interval models if the delay length is long. However, one unexpected event encountered in the current study was our high exclusion rate, particularly in Experiment 1 (15 out of 63 participants), which is higher than McAuley and Jones' original Experiment 4 (0 exclusions reported out of 14 participants). Mostly, we excluded the participants due to low accuracy (10 out of 15 participants) on the most distinguishable (easiest) comparison intervals in one or more blocks of test trials. Also, those 10 participants tended to have less music training compared to participants included in the analysis (2.5 years vs. 6.9 years). Thus, one explanation for the difference between the current study and McAuley and Jones' original experiment might be the music training: while we had a sample with diverse music experiences, it is possible that only high-experience individuals volunteered for McAuley and Jones' study and that experience (or preexisting differences in ability) led those participants to perform more accurately. Nonetheless, it is clear that there is a wide range of performance in this task, and it is possible that different participants are using different mechanisms.

More broadly, explaining differences among individuals is an important issue faced in the field. Several studies have shown individual differences in both behavioral and neural outcomes. For example, Matthews (2011, Experiment 2) found different patterns across individuals in a duration discrimination task (those results are replotted in Matthews and Meck's 2014 review paper Figure 3). They concluded that people perform quite differently in the behavioral tasks and this is likely because they activated different neural temporal processing mechanisms when doing the tasks. Grahn and McAuley (2009) also investigated people's brain activations as a function of their natural tendencies to perceive an implied beat or not. They found that the activation in auditory and motor areas was correlated with individual differences in perceiving beats, even though the behavioral performances were similar. This again reminds us that individual differences are one of the major issues requiring more explorations in the field of time perception (Matthews & Meck, 2014), in terms of its behavioral, experiential, and neural bases.

Future directions.

Two important future directions are exploring how entrainment decay interacts with beat production, and exploring the neural underpinnings of entrainment decay. The modified M & K paradigm used in the current study is a relatively passive behavioral measure (i.e. listening paradigm) compared to other paradigms in which listeners actively generate a metrical response (e.g. finger-tapping). Future research might use active behavioral measures as well as neural measures to enable a more direct investigation of the entrainment process. For example, recording active rhythm production measures, such as synchronization error (see Repp, 2005 and Repp & Su, 2013 for a thorough review of tapping experiments), might allow researchers to observe a more precise timing of entrainment decay across time. Further, this might highlight differences in entrainment duration depending on the level of cognitive engagement. The entrainment effect might last longer with active body engagements from an embodiment perspective.

In a different direction, neural imaging tools such as EEG/MEG or fMRI and TMS can be used to characterize the decay of entrainment effect in terms of brain activity, potentially clarifying the neural mechanisms underlying time perception. Based on fMRI data, researchers have posited that different brain networks may compute duration perception. One specific proposal is that the cerebellum circuit is implicated in interval timing and the basal ganglia circuit is more involved in entrainment timing (Buhusi & Meck, 2005; Grondin, 2010, Teki et al., 2011). The basal ganglia circuit may be less functionally specific than the cerebellar circuit, with some findings of basal ganglia supporting a role in encoding of time interval information (Harrington et al., 2004; Rao et al., 2001). Only a few recent studies have directly contrasted entrainment-like and interval-like duration perception systems. Recently, Breska and Ivry (2018) showed double dissociation evidence for two timing systems: individuals with cerebellar degeneration showed deficits in a single-interval cueing task (similar to interval timing), but not a rhythm cueing task (similar to entrainment), while those with basal ganglia dysfunction showed the reverse pattern of difficulty. Moreover, duration perception has also been linked to cortical networks such as premotor cortex, prefrontal cortex, parietal cortex (Bueti et al., 2008; Harrington, Haaland, & Knight, 1998); modality-specific areas such as V5/MT for visual temporal processing (Bueti, Bahrami, & Walsh, 2008); and supramodal auditory cortex for time estimation of both auditory and visual stimuli (Kanai et al., 2011). See Ivry and Schlerf (2008) and Grondin (2019) for an overview of different underlying neural mechanisms of time perception.

The behavioral results from our modified M & K paradigm in the current study suggest a differential activation of interval and entrainment circuits as a function of delay length. Our view is that duration judgement performance can be explained by the weighted combination of the interval and the entrainment models (and, at a mechanistic level, brain circuits), depending on how far the entrainment effect has decayed. One interesting and important test of this hypothesis is whether there is evidence for a tradeoff in neural systems in play as entrainment decays, which might be fruitfully investigated using MEG, which has both temporal precision and high spatial resolution.

Conclusion

In a modified duration discrimination paradigm, we found a duration judgement pattern similar to the prediction of the entrainment models when there was a short delay between standard and comparison intervals, and a pattern similar to the predictions of interval models when there were longer delays between standard and comparison intervals. Our data suggest that the entrainment effect may disappear somewhere between 2 beats and 4 beats according to the entraining context, and thus, suggests the importance of adding a "delay parameter" to make the entrainment models more complete. This parameter should take varied delay lengths (i.e. the number of beats) into account to conduct the period and phase correction, and finally calculate the temporal contrast. More importantly, we found that accuracy is not impaired when entrainment decays--in fact, it increases. This finding suggests an interplay between entrainment and interval timings as a function of delay length.

References

Allan, L. G. (1979). The perception of time. *Perception & Psychophysics*, *26*(5), 340-354.

Allan, L. G. (1998). The influence of the scalar timing model on human timing research. *Behavioural Processes*, *44*, 101-117.

Bueti, D., Bahrami, B., & Walsh, V. (2008). Sensory and association cortex in time perception. *Journal of Cognitive Neuroscience*, *20*(6), 1054-1062.

Bueti, D., Walsh, V., Frith, C., & Rees, G. (2008). Different brain circuits underlie motor and perceptual representations of temporal intervals. *Journal of cognitive neuroscience*, *20*(2), 204-214.

Brainard, D. H. (1997) The Psychophysics Toolbox, *Spatial Vision, 10*, 433-436.

Breska, A., & Ivry, R. B. (2018). Double dissociation of single-interval and rhythmic temporal prediction in cerebellar degeneration and Parkinson's disease. *Proceedings of the National Academy of Sciences*, *115*(48), 12283-12288.

Buhusi, C. V., & Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews Neuroscience*, *6*(10), 755.

Campbell, W. N., & Isard, S. D. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, *19*(1), 37-47.

Church, R. M., Meck, W. H., & Gibbon, J. (1994). Application of scalar timing theory to individual trials. *Journal of Experimental Psychology: Animal Behavior Processes, 20*(2), 135–155.

Calderone, D. J., Lakatos, P., Butler, P. D., & Castellanos, F. X. (2014). Entrainment of neural oscillations as a modifiable substrate of attention. *Trends in cognitive sciences, 18*(6), 300-309.

Carrasco, M. C., Bernal, M. C., & Redolat, R. (2001). Time estimation and aging: a comparison between young and elderly adults. *The International Journal of Aging and Human*

Development, *52*(2), 91-101.

Coelho, M., Ferreira, J. J., Dias, B., Sampaio, C., Martins, I. P., & Castro-Caldas, A. (2004). Assessment of time perception: The effect of aging. *Journal of the International Neuropsychological Society*, *10*(3), 332-341.

Coull, J. T., Cheng, R. K., & Meck, W. H. (2011). Neuroanatomical and neurochemical substrates of timing. *Neuropsychopharmacology*, *36*(1), 3.

Cowan, N., Saults, J. S., & Nugent, L. D. (1997). The role of absolute and relative amounts of time in forgetting within immediate memory: The case of tone-pitch comparisons. *Psychonomic Bulletin & Review, 4*(3), 393-397.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Syllabic stress. *The Journal of the Acoustical Society of America*, *83*(4), 1574-1585.

Drake, C., & Botte, M. C. (1993). Tempo sensitivity in auditory sequences: Evidence for a multiple-look model. *Perception & Psychophysics, 54*(3), 277-286.

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

Franz, V. H., & Loftus, G. R. (2012). Standard errors and confidence intervals in within-subjects designs: Generalizing Loftus and Masson (1994) and avoiding the biases of alternative accounts. *Psychonomic bulletin & review*, *19*(3), 395-404.

Gibbon, J. (1971). Scalar timing and semi-Markov chains in free-operant avoidance. *Journal of Mathematical Psychology*, *8*, 109-138.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review, 84*, 279–325.

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. *Annals of the New York Academy of Sciences, 423*, 52–77.

Grahn, J. A., & McAuley, J. D. (2009). Neural bases of individual differences in beat perception. *NeuroImage, 47*(4), 1894-1903.

Grondin, S. (2010). Timing and time perception: a review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics*, *72*(3), 561-582.

Grondin, S. (2020). *The Perception of Time: Your Questions Answered*. Routledge.

Grondin, S., & Rammsayer, T. (2003). Variable foreperiods and temporal discrimination. *The Quarterly Journal of Experimental Psychology Section A*, *56*(4), 731-765.

Harrington, D. L., Boyd, L. A., Mayer, A. R., Sheltraw, D. M., Lee, R. R., Huang, M., & Rao, S. M. (2004). Neural representation of interval encoding and decision making. *Cognitive Brain Research*, *21*(2), 193-205.

Harrington, D. L., Haaland, K. Y., & Knight, R. T. (1998). Cortical networks underlying mechanisms of time perception. *Journal of Neuroscience*, *18*(3), 1085-1095.

Honing, H. (2003). The final ritard: On music, motion, and kinematic models. *Computer Music Journal*, *27*(3), 66-72.

Ivry, R. B., & Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends in cognitive sciences*, *12*(7), 273-280.

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review, 83*, 323–335.

Kanai, R., Lloyd, H., Bueti, D., & Walsh, V. (2011). Modality-independent role of the primary auditory cortex in time estimation. *Experimental Brain Research*, *209*(3), 465-471.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208-1221.

Lakatos, P., Gross, J., & Thut, G. (2019). A new unifying account of the roles of neuronal entrainment. *Current Biology*, *29*(18), R890-R905.

Large, E. W., Herrera, J. A., & Velasco, M. J. (2015). Neural networks for beat perception in musical rhythm. *Frontiers in systems neuroscience, 9*, 159.

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review, 106*(1), 119.

Large, E. W., and Snyder, J. S. (2009). Pulse and meter as neural resonance. *Annals of the New York Academy of Sciences, 1169*, 46–57.

Li, Q., Liu, G., Wei, D., Liu, Y., Yuan, G., & Wang, G. (2019). Distinct neuronal entrainment to beat and meter: Revealed by simultaneous EEG-fMRI. *NeuroImage, 194*, 128-135.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

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*(5), e19815.

Matthews, W. J., & Meck, W. H. (2014). Time perception: the bad news and the good. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(4), 429-446.

Mathewson, K. E., Fabiani, M., Gratton, G., Beck, D. M., & Lleras, A. (2010). Rescuing stimuli from invisibility: Inducing a momentary release from visual masking with pre-target entrainment. *Cognition*, *115*(1), 186-191.

Mathewson, K. E., Gratton, G., Fabiani, M., Beck, D. M., & Ro, T. (2009). To see or not to see: prestimulus α phase predicts visual awareness. *Journal of Neuroscience*, *29*(9), 2725-2732.

McAuley, J. D., & Jones, M. R. (2003). Modeling effects of rhythmic context on perceived duration: a comparison of interval and entrainment approaches to short-interval timing. *Journal of Experimental Psychology: Human Perception and Performance, 29*(6), 1102.

McAuley, J. D., & Kidd, G. R. (1998). Effect of deviations from temporal expectations on tempo discrimination of isochronous tone sequences. *Journal of Experimental Psychology: Human Perception and Performance, 24*(6), 1786.

Nozaradan, S., Peretz, I., Missal, M., Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *Journal of Neuroscience, 31*, 10234–10240.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies, *Spatial Vision, 10*, 437-442.

Rakitin, B. C., Gibbon, J., Penney, T. B., Malapani, C., Hinton, S. C., & Meck, W. H. (1998). Scalar expectancy theory and peak-interval timing in humans. *Journal of Experimental Psychology: Animal Behavior Processes, 24*(1), 15–33.

Rao, S. M., Mayer, A. R., & Harrington, D. L. (2001). The evolution of brain activation during temporal processing. *Nature neuroscience*, *4*(3), 317.

Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei". *The Journal of the Acoustical Society of America, 92*(5), 2546-2568.

Repp, B. H. (1998a). The detectability of local deviations from a typical expressive timing pattern. *Music Perception: An Interdisciplinary Journal, 15*(3), 265-289.

Repp, B. H. (1998b). Variations on a theme by Chopin: Relations between perception and production of timing in music. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 791.

Repp, B. H. (2002). Phase correction in sensorimotor synchronization: Nonlinearities in voluntary and involuntary responses to perturbations. *Human Movement Science, 21*(1), 1-37.

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychonomic bulletin & review, 12*(6), 969-992.

Repp, B. H., & Su, Y. H. (2013). Sensorimotor synchronization: a review of recent research (2006–

2012). *Psychonomic bulletin & review*, *20*(3), 403-452.

Snyder, J. S., Carter, O. L., Lee, S.-K., Hannon, E. E., & Alain, C. (2008). Effects of context on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 1007–1016.

Snyder, J. S., & Weintraub, D. M. (2013). Loss and persistence of implicit memory for sound: Evidence from auditory stream segregation context effects. *Attention, Perception, & Psychophysics, 75*(5), 1059-1074.

Teki, S., Grube, M., Kumar, S., & Griffiths, T. D. (2011). Distinct neural substrates of duration-based and beat-based auditory timing. *Journal of Neuroscience*, *31*(10), 3805-3812.

Umeda, N. (1975). Vowel duration in American English. *The Journal of the Acoustical Society of America, 58*(2), 434–445.

Appendix

In this paragraph, we will review the essential formulas in McAuley and Jones' (2003) framework and further explain how they developed the predictions from the two models.

Based on the framework of McAuley and Jones (2003), a timekeeper (which can be regarded as the internal clock in interval models or the oscillator in entrainment models) conducts linear phase and period corrections by calculating the temporal contrast. This framework is useful because it can be used to generate predictions aligned with both interval models and entrainment models by changing two parameters. They proposed the following three formulas:

$$(1)\ \ if\ \ (\phi_{i-1} + \frac{IOI_i}{P_i}) > 0.5,\ \ C = (\phi_i + \frac{IOI_i}{P_i})(\mathrm{mod}\ \ 1) - 1$$

$$otherwise,\ \ C = (\phi_i + \frac{IOI_i}{P_i})(\mathrm{mod}\ \ 1)$$

$$(2)\ \ \phi_{i+1} = (1 - W_\phi)C(\phi_i, IOI_i, P_i)$$

$$(3)\ \ P_{i+1} = [1 + W_p C(\phi_i, IOI_i, P_i)]P_i$$

C represents the temporal contrast corresponding to the *i*th stimulus. It is corrected to the range between -0.5 to 0.5 in their formula, as determined by ϕ_i, IOI_{i}, and P_i. ϕ_i is the relative phase of the *i*th stimulus onset. IOI_i is the interval between the onset of the *i*th stimulus and the onset of the next stimulus. P_i is the period of the current oscillator. The default of P_i is set to the first IOI. McAuley and Jones set the first ϕ to zero, which means the phase of the timekeeper completely matches the onset of the first external stimulus. W_ϕ determines the extent of phase correction. This parameter varies from 0 to 1. When W_ϕ equals zero, the timekeeper does not modify the phase to follow the onset of the external stimuli. When W_ϕ is equal to one, the timekeeper updates the phase in every cycle to completely match the onset of the external stimuli. W_{P} determines the extent of the period reset. Like W_ϕ, this parameter also varied from 0 to 1. When W_{P} is equal to zero, the timekeeper does not change the period to match the change of the IOI. When W_{P} is equal to one, the timekeeper resets its period to match each IOI.

McAuley and Jones' (2003) framework can be used to generate predictions aligned with both interval models and entrainment models by changing W_ϕ and W_P. As in the full phase reset hypothesis, the timekeeper with W_ϕ=1 mimics the characteristic of interval models. On the other hand, the entrainment models have a flexible timekeeper with varied values of parameters W_ϕ and W_P. In McAuley and Jones' (2003) Experiment 4, they estimated the parameters for the entrainment model to be $W_\phi = 0.8$, $W_P = 0.05$ and the interval model to be $W_\phi = 1$, $W_P = 0$. By using their parameters for our study, which is very similar to McAuley and Jones' (2003) Experiment 4, we expected similar model predictions. By this setting of parameters and the manipulation of the comparison onset times, which provide the IOI for each stimulus, we can calculate the temporal contrast (C). By assuming that response probabilities approximated a normal distribution, $C < 0$ is converted to responding *shorter* and $C > 0$ to responding *longer*. Then we can calculate the predicted proportion of responding short, and then use that quantity to predict PSE for each model. The interval models predicted the proportion of responding short or long to correspond perfectly to whether the comparison intervals are actually shorter or longer than the standard interval, because the full phase reset prevents the timekeeper from being influenced by the jitter of the delay length. Accordingly, it predicts a PSE of 600 ms regardless of the comparison onset time manipulation. The prediction of entrainment models is quite different. With a manipulation of unexpectedly early trials (i.e. -180 ms onset time), the entrainment model predicted 100% of responding short in the first four comparison intervals (i.e. 550, 570, 590, 610 ms), and 0% of responding short in the last two comparison intervals (i.e. 630, 650ms) due to the compression of the delay length of 180 ms. Therefore, the proportion short response would be 0.67 in the early trials. With a manipulation of on-time trials, the entrainment model predicted 100% of responding short in the first three comparison intervals (i.e. 550, 570, 590 ms) and 0% of responding short in the last three comparison intervals (i.e. 610, 630, 650ms), and thus the proportion short response would be 0.5.

With a manipulation of unexpectedly late onset trials, the entrainment model predicted 100% of responding short in the first two comparison intervals (i.e. 550, 570 ms) and 0% of responding short in the last four comparison intervals (i.e. 590, 610, 630, 650ms), thus the proportion short response would be 0.33. In sum, with $W_{\phi} = 0.8$ and $W_P = 0.05$, the estimated proportion short of early trials equals 0.67 and estimated PSE of early trials ranges between 610 ms and 630 ms. The estimated proportion short of ontime trials equals 0.5 and the estimated PSE of on-time trials ranges between 590 ms and 610 ms. The estimated proportion short of late trials equals 0.33 and the estimated PSE of late trials ranges between 570 ms and 590 ms (Table 1). Please note that changing the W_{ϕ} and W_P yields different estimated proportion short responses (i.e. 0, 0.17, 0.33, 0.5, 0.67, 0.83, 1) and the corresponding PSEs.

As noted in the main manuscript, in Experiment 1, our short-delay results fit M&J’s entrainment-model parameters better than their interval-model parameters, while our long-delay results fit the interval-model parameters. We considered refitting the model to our data. Without knowing exactly how M&J arrived at their parameter values, we explored the 2-dimensional space of multiple parameter value combinations in early, ontime and late conditions (Figure A1). As it turns out, a range of parameter values can predict our outcomes (and theirs, please see their Figure 5). To achieve either entrainment-like or interval-like outcomes corresponding to different comparison onsets, the phase weight and period weight essentially trade off. On-time trials are not informative with respect to cue weights since proportion short is identical regardless of cue weights. We overlaid the left (early trials) and right (late trials) figures and found that only small regions fit our proportion short results, briefly summarized as below. EXP1 short-delay condition (entrainment-like results, highlighted in yellow): 0.67 (~0.67), 0.50 (~0.50), 0.45 (~0.50) corresponding to early, ontime and late trials; EXP2 short-delay condition (entrainment-like results, highlighted in orange): 0.81 (~0.83), 0.45 (~0.50), 0.35 (~0.33) corresponding to early, ontime and late trials; Long-delay conditions (interval-like results, highlighted in purple), including EXP1 long-delay, EXP2 medium-delay and EXP2 long-delay conditions: 0.54, 0.52, 0.52, 0.53, 0.49, 0.47, 0.49, 0.48, 0.48 (all ~0.5) corresponding to early, ontime and late trials, respectively. Please find the highlighted regions in Figure A2.

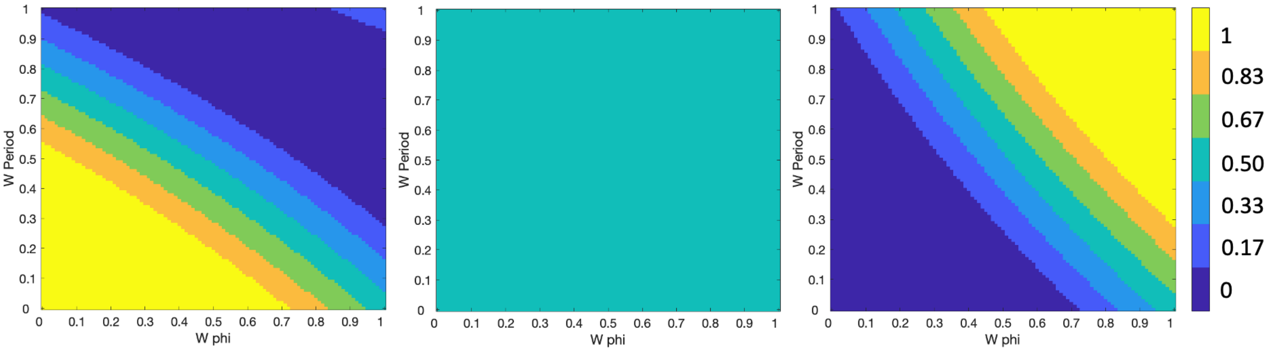

Figure A1. Model prediction of proportion short response in early (left), ontime (middle) and late (right) trials with different combinations of phase and period parameters.

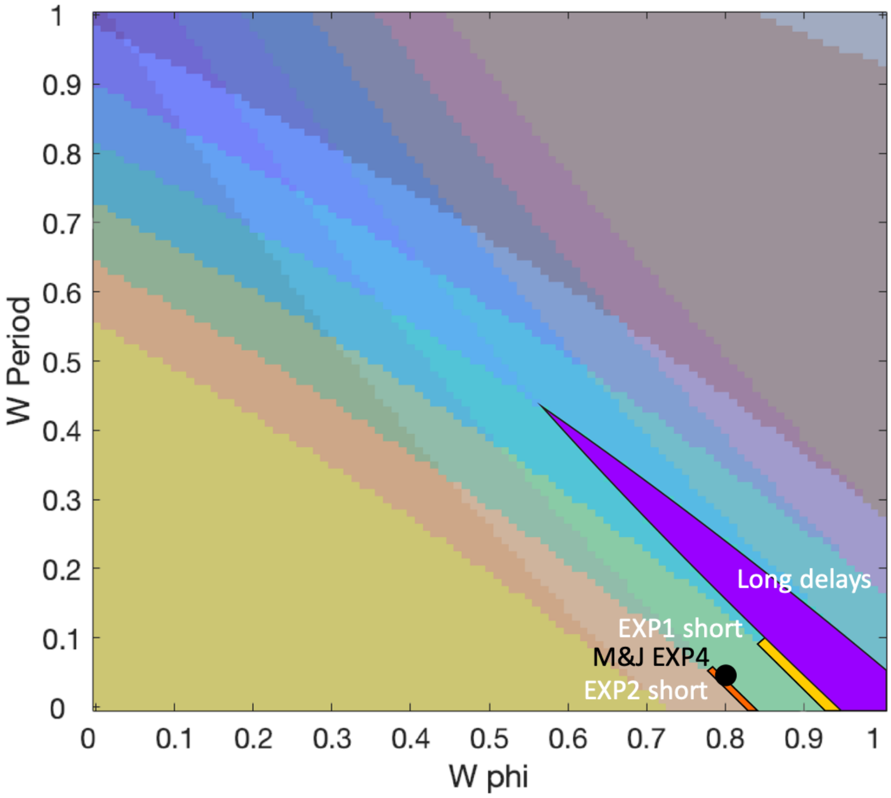

Figure A2. Highlighted regions that fit our proportion short results. The region highlighted in purple represents the Experiment 1 long-delay, Experiment 2 medium-delay and Experiment 2 long-delay conditions. The region highlighted in yellow represents the Experiment 1 short-delay condition. The region highlighted in orange represents the Experiment 2 short-delay condition. The black dot represents the entrainment-model parameters ($W_\phi = 0.8$, $W_P = 0.05$) used in McAuley and Jones' Experiment 4 (2003).