Humans don't time sub-second intervals like a stopwatch

Marta Narkiewicz, Anna Lambrechts, Frederik Eichelbaum & Kielan Yarrow*

Department of Psychology,

City University London

* Author for correspondence:

Kielan Yarrow,
Social Science Building,
City University,
Northampton Square,
London EC1V 0HB

Tel: +44 (0)20 7040 8530
Fax: +44 (0)20 7040 8580
Email: kielan.yarrow.1@city.ac.uk

Abstract

Many activities require the ability to estimate intervals of time in an accurate and flexible manner. A traditional and popular account suggests that humans possess a kind of internal stopwatch that can be started, paused and stopped at will. Here we test this idea by measuring variable performance errors in three experiments. Participants had to compare the total time accumulated during one to three short target intervals with a single standard interval. With two or more target intervals, participants had to pause, but not reset, their putative internal stopwatches. By establishing baseline performance at two different standard durations and extrapolating based on Weber's law, we were able to estimate how much performance should have deteriorated when target segments contained breaks. The decrement in performance we observed far exceeded the stopwatch prediction, and also exceeded the simulated predictions of a modified stopwatch with a slowing pacemaker. The data thus favour either a counter that cannot be paused during sub-second durations or alternative models of sub-second interval duration discrimination which do not posit a count-based metric for time. We discuss several possible strategies which participants might have implemented in order to apply such clocks in the split-interval task.

Keywords: Time perception; counter; pacemaker-accumulator; switch; intervals.

Introduction

Interval timing underpins a wide range of sensory, cognitive and motor behaviours, so it is not surprising that humans and other animals are able to attain relatively accurate and precise timing in the milliseconds to minutes range (Allan, 1979; Grondin, 2001). Many models have been proposed in order to illustrate and explain timing performance, suggesting various metaphors for the timing process. These range from binary oscillators to plastic neural networks (Matell & Meck, 2004). However, the most pervasive metaphor is that of the stopwatch.

In their most basic form, such counter models (Creelman, 1962; Treisman, 1963) posit a pacemaker which generates a series of pulses that measure out roughly equal units of time. These are integrated, for example by being sent to an accumulator and stored, whilst a "switch" is closed, but are no longer accumulated when the switch is open. The switch itself would typically be closed only during an interval of interest that is being timed. The collected pulses then correspond to the amount of time which has passed, so that when an interval comparison is required, the accumulated pulses can be compared with a stored value representing a specific duration. Despite the rise of alternate timing models in recent years (for reviews see e.g. Buhusi & Meck, 2005; Ivry & Schlerf, 2008), the counter model remains prominent in the field of temporal research, most likely due to its conceptual simplicity, and also its flexibility, which allows it to explain a wide range of timing behaviours.

Testing stopwatch models with a (broken) interval comparison task

An established method in prospective timing research involves discriminating between two different intervals – a standard, which either remains constant throughout the experiment or "roves" between a small number of base durations, and a target interval which varies from trial to trial,

straddling the standard/s (Grondin, 2001). Counter models are well able to predict behaviour in this task. Critically, in this paper we additionally split the target interval into two or more parts and ask participants to combine these durations and then compare their *sum* to the standard. Similar experiments, albeit with longer intervals and somewhat different tasks, have been conducted in both animals and humans (e.g. Buhusi, Sasaki, & Meck, 2002; Buhusi & Meck, 2006; Fortin & Masse, 2000) although the use of longer intervals would allow for counting strategies (in humans), which is an issue we avoid here by using sub-second intervals.[1]

On our reading, according to popular counter models such as Scalar Expectancy Theory (SET), all that would be required when timing a broken target interval (compared to an unbroken one) is an additional opening and closing of the switch (to halt accumulation during the break). The accumulated count should be maintained without difficulty while the switch is closed. This is the key feature of pacemaker-accumulator models that we test in this paper, so it bears repeating: We will test the idea that a pacemaker-accumulator internal clock can continue to accumulate following pauses in accumulation, i.e. that opening the switch does not mandatorily reset the accumulator (so that counter models can be considered as *pausable* stopwatches).

We believe that this "pause" feature is implicit in most counter models. Indeed, theorists have previously argued that this is possible using the SET framework (e.g. Lejeune, 1998; Roberts & Church, 1978), and have used such a pause/restart process to explain success on split-interval tasks in humans and animals. For example, Fortin and Masse (2000) had participants reproduce a target duration after hearing an interval containing a break, and participants were found to reproduce intervals fairly accurately. The fact that participants were successfully taking account of the breaks seems to support the idea of counter-based clock that can be paused, and was interpreted in this

[1] Note that while some counter models were not initially intended to deal with short intervals, counter models have often been posited to account for experimental results with stimuli as short as 100 ms (e.g. Getty, 1975; Rammsayer & Ulrich 2005; Wearden, Edwards, Fakhri, & Percival, 1998).

context (although this was not the main subject of investigation). Furthermore, these authors and their collaborators have subsequently shown that accurate timing can be obtained from humans with broken intervals in several tasks, not just reproduction (Fortin & Tremblay, 2006; Fortin et al., 2009; Tremblay & Fortin, 2003). In a similar vein, animal work suggesting success at pausing the timer during gaps in the peak-interval procedure (e.g. Roberts & Church, 1978; see discussion for further details) appears to have partly motivated the original inclusion of the switch within the SET framework (Buhusi & Meck, 2000).

In the interests of balance, we should note that a variety of counter models have been described, and some authors may have intended that switch and accumulator reset operations should be considered mechanically linked, such that opening the switch automatically resets the accumulator. However, we are not aware of any unambiguous statements to this effect, so we consider a test of the ability to pause timing to be a reasonable test of most counter models, as currently described.

Stopwatch predictions about precision

In temporal psychophysics, it is standard practice to parcel up trial-by-trial errors into bias, or constant error, which reflects mean accuracy over multiple trials, and variable error (or its inverse, precision) which reflects the extent of trial-by-trial deviation from the average. The apparent successes at pausing and restarting accumulation described above stem from analyses focussing on accuracy. Here, we instead focus on predictions about precision, such that the logic we present below has not, to our knowledge, been expounded or tested before.

Having to open and close the switch one more time to pause accumulation could add noise to the task (arising, for example, from variability in the latencies with which sensory signals reach the switch). But how much extra noise should we expect? To answer this question, we can turn to the

known psychophysical properties of time perception. Humans generally show a linear relationship between interval duration and the standard deviation of the resulting trial-to-trial noise (Wearden & Lejeune, 2008). This is known as the scalar property. The trial-to-trial noise that is proportional to stimulus duration is not, however, the only source of variance in timing tasks. A noise component that arises irrespective of stimulus magnitude is also ubiquitous, such that the variability obtained on a typical duration discrimination task is often considered to stem from both sources added together (Getty, 1975). This is most evident for very short intervals, where the small constant component of the variable error is not swamped by a large scalar component. Hence the overall precision of interval timing is probably best described as following a generalised version of Weber's law (Wearden & Lejeune, 2008), although the exact details of this relationship continue to be debated (e.g. Bangert, Reuter-Lorenz, & Seidler, 2011; Crystal, 1999; Lewis & Miall, 2009; Matthews & Grondin, 2012).

The variance observed in an interval discrimination task can therefore be conceived of as an additive combination of non-scalar variance (independent of duration) and variance that scales with interval duration (scalar variance). The key intuition required to understand our initial predictions here is that operations like opening and closing the switch contribute only to the *non-scalar* variance (because for a typical timing task, they occur in a similar manner regardless of interval duration). Critically, we can make use of the "slope" analysis method (Ivry & Hazeltine, 1995) in order to estimate the non-scalar variance separately from the total variance. In our experiment, it is expected that performance will deteriorate in the split-interval task when compared to a classic comparison task, but only by an amount predictable from an estimate of the non-scalar variance, as outlined next.

The maximum variance associated with the additional opening and closing of the switch can be calculated using the generalised version of Weber's Law. The standard deviation of variable errors increases linearly with interval duration, which means that total variance follows a power function:

$$\sigma^2_{observed} = st^2 + c \qquad (1)$$

Where *t* is the standard interval's duration, *s* is the scalar variance and *c* is the constant (non-scalar) variance. An estimate of *s* and *c* can then be obtained by assessing performance at two or more standard intervals and extrapolating the function that joins these points trough a notional interval duration of zero (i.e. performing a non-linear regression and seeking the intercept; see Figure 2 panel b for an illustration).

In the classic interval comparison task, the switch will be opened and closed an equal number of times irrespective of interval duration, and therefore the variance associated with these two switch operations (σ^2_{switch}) is non-scalar and included in *c:*

$$\sigma^2_{switch} \leq c \qquad (2)$$

Considering the classic interval comparison task further, one approach that might be employed by the participant when facing this task would involve two repetitions of the combined close and open operation – one to estimate the target interval and one to estimate the standard interval:

$$\sigma^2_{switch} = 2.\ \sigma^2_{close\&open} \qquad (3)$$

Under this same strategy, the broken-interval task would require both switch operations from the classic task, together with an additional switch operation for the second segment of the (broken) target interval. In other respects (i.e. counter, memory and decision processes) the task is formally

identical to the classic interval comparison task. Therefore, performance on the broken-interval task should deteriorate ($\sigma^2_{increase}$) by no more than:

$$\sigma^2_{Increase} \leq c/2 \qquad (4)$$

It is also possible that participants adopt another strategy whereby they form an internalised standard (based on earlier trials from the experiment) and thus do not rely on a new estimate for the standard interval on every trial. In this case, for the classic task, only one close and open operation would generally be carried out per trial (for the target interval) so σ^2_{switch} is then an estimate of just one switch operation:[2]

$$\sigma^2_{switch} = \sigma^2_{close\&open} \qquad (5)$$

In this instance the broken-interval task would require the single switch operation from the classic task, together with an additional switch operation for the second segment of the (broken) target interval. Consequently, performance should not deteriorate in the split-interval task by more than *c* compared to the standard interval comparison task:

$$\sigma^2_{Increase} \leq c \qquad (6)$$

This second prediction was adopted in our experiments as it gives stopwatch models more leeway for success. A total of three experiments were conducted. The first two specifically focussed on assessing how much performance deteriorated in the split-interval task in order to test constant-rate

[2] This method would actually be a superior strategy as it would eliminate the variance associated with measuring the standard on each trial (Morgan, Watamaniuk, & McKee, 2000) although it might be more demanding, particularly in the present set of experiments, which include two interleaved standards (see e.g. Acerbi, Wolpert & Vijayakumar, 2012, Taatgen & Van Rijn, 2011, for the use of two standards / bimodal priors in interval timing)

pacemaker models like SET at short intervals. The last experiment simultaneously examined the more complex predictions derived from a modified counter module, proposed by Taatgen et al. (2007), who suggest that pulses become more distributed as the interval duration increases. These latter predictions are outlined in more detail later in the article, prior to describing the relevant experiments. In general, we find that our data from sub-second intervals depart substantially from the predictions of a stopwatch that can be paused.

Experiment 1

Methods

Participants

Nine women and three men were paid to participate. Two participants were outliers (> 2xSD from the group mean) on the two key dependent variables (see data analysis) and were replaced by two of the authors[3] to yield a final sample of eight women and four men (mean age = 24.7, SD = 7.2).

Apparatus and Stimuli

The experiment was controlled by a PC sending digitised signals at 44100 Hz using a 12 bit A/D card (National Instruments DAQ Card 6715). We confirmed the correct timing of output signals using a 20 MHz storage oscilloscope (Gould DSO 1604). Stimuli were 1000 Hz pure tones of various durations, presented via a small speaker placed in front of the participant.

[3] Note that inclusion of these two excluded participants would actually have increased the critical difference we observed ($\sigma^2_{Increase}$ vs. c), albeit at the cost of greater variability: One case involved a large negative estimate for the non-scalar variance, and the other involved a large decrement in broken-interval conditions compared to single interval conditions. A wilcoxon test on the initial sample was highly significant for our key comparison ($p = 0.002$).

<INSERT FIGURE 1 AROUND HERE>

Design and procedure

A 2x3 repeated-measures design included two factors: *standard duration* (300 and 600 ms) and *target stimulus* (single, double, triple). The two standard durations were randomly interleaved within a block of 120 trials. Each trial commenced with the presentation of a target segment comprising one to three comparison stimuli, with the number of target stimuli held constant in a block, and block order counterbalanced across participants. Where more than one target stimulus was presented, they were separated by 500 ms breaks. After the target segment came a 1000 ms silent period and then the standard stimulus. Participants judged whether the total time for which the tone was on during the target segment was longer or shorter than the standard stimulus (see Figure 1a). Total on time for target stimuli was selected at random on each trial from an adaptive condition-specific distribution (Rosenberger & Grill, 1997), a procedure intended to sample a sufficient range of values to capture the psychometric function without wasting trials at the extremes. The distribution was initially uniform (70-130% of standard duration in 5% steps) but could expand towards 25% and 175% limits based on the responses participants made. When the target segment contained more than one stimulus, total on time was divided evenly between the different stimuli.

Data Analysis

In each condition, a maximum-likelihood cumulative Gaussian fit was obtained to the proportion of times that the target was judged longer than the standard for each tested target duration, using the Psignifit toolbox (Wichmann & Hill, 2001) in Matlab (MathWorks, Natick, MA). To measure variable error (i.e. precision) we estimated the standard deviation of the cumulative Gaussian ($\sigma_{Observed}$) using

the difference between durations required to yield "long" judgements 84% and 50% of the time. We then converted this value to judgement variance ($\sigma^2_{Observed}$) by squaring. We also estimated the point of subjective equality (PSE, i.e. constant error or accuracy) from the 50% point of the fitted function. The $\sigma^2_{Observed}$ was used to estimate how much performance should have deteriorated ($\sigma^2_{Increase}$) when the internal clock had to be started and stopped one more time. $\sigma^2_{Increase}$ was calculated by taking the average of four changes in $\sigma^2_{Observed}$: One each when going from single to double and from double to triple conditions, at standard durations of both 300 and 600ms. Our key prediction involved comparing this value with an estimate of the non-scalar variance (*c*). The scalar property states that scalar variance increases linearly with the square of the standard duration. Hence non-scalar variance was found by determining where a power function fitting the two relevant $\sigma^2_{Observed}$ points (the 300 and 600 ms single target stimulus conditions) crossed the line x = 0 (i.e. the y axis; see Figure 2b).[4]

We investigated possible differences across conditions using repeated-measures t-tests and/or factorial repeated-measures ANOVA, with Greenhouse-Geisser corrections applied to control any violations of sphericity and alpha set at 0.05. Although our key hypothesis was directional, all reported p values are two-tailed.

Results and Discussion

Figure 2 panels A-E shows data from our first experiment. Panel A shows the combined data from all participants in the 600 ms standard conditions, along with cumulative Gaussian fits, in order to illustrate the fitting procedure. Notice how the slopes of the fitted functions are flatter in the double

[4] In order to obtain an approximately unbiased estimator for non-scalar variance (in the formal statistical sense), for a fair comparison with $\sigma^2_{Increase}$, we retained individual estimates < 0. However, replacing such estimates with zero made no difference to the patterns of statistical significance reported in Experiments 1 & 3. The exception to this was Experiment 2; see footnote #7 for details.

and triple conditions compared to the single condition, implying a loss of precision when the target interval was broken up into parts. For our actual analysis, separate fits were obtained for each participant, and used to estimate judgement variance ($\sigma^2_{Observed}$). Figure 2 panel B shows values of $\sigma^2_{Observed}$ for each of the 12 participants in the single-target (i.e. classic/baseline task) conditions, along with the power function drawn through the estimates obtained with standard durations of 300 and 600 ms. We obtained an estimate of non-scalar variance (*c*) for each participant based on the point at which these functions crossed the y axis.

<INSERT FIGURE 2 AROUND HERE>

Figure 2 part C shows mean values of $\sigma^2_{Observed}$ in all six conditions of Experiment 1. Judgement variance increased from the single to the double to the triple conditions ($F_{[1.2,13.1]} = 7.9$, $p = 0.012$; all protected LSD follow-ups $p < 0.05$) with no other significant effects. Although not central to our analysis, Figure 2 part D shows points of subjective equality (shown as a bias, i.e. with the standard duration subtracted) matched to the values of $\sigma^2_{Observed}$ shown in part C. Here, standard duration interacted with the number of target stimuli ($F_{[1.2,13.6]} = 5.1$, $p = 0.036$) but there were no significant differences across number of target stimuli for either 300 ms or 600 ms standards.

To test the stopwatch model of interval timing, we derived an estimate of nonscalar variance from the single-target conditions (*c* from Equation 1), and compared it to the average decrement in performance observed when the putative internal stopwatch had to be started and stopped one more time ($\sigma^2_{Increase}$). This key comparison is shown in Figure 2 part E. The mean value of $\sigma^2_{Increase}$ far exceeded the mean value of *c* ($t_{[11]} = 3.3$, $p= 0.007$). This finding demonstrates clearly that when changing from a task where one interval must be compared with a standard, to a task where the internal clock must be paused one or more times during the process of accumulation before comparison with a standard, performance deteriorates dramatically. This deterioration cannot be

accounted for by the variance that is introduced every time a switch which gates pulses into an accumulator is used to start and stop the clock.

Experiment 2

Although Experiment 1 showed a far greater decrement for sub-second broken interval tasks than stopwatch models would predict, we were concerned about two possible objections to our method. Firstly, variance might be introduced into temporal judgements whenever temporal estimates are stored in memory. Because we equated the time from the end of the target stimuli to the start of the standard stimuli in Experiment 1, the overall length of trials was shortest for single-target conditions, slightly longer for double conditions, and longer still for triple conditions (see Figure 1a). To address this issue our second experiment matched overall trial duration between single and double conditions, while dropping the superfluous triple conditions (Figure 1b).

Secondly, we wondered whether the presumably quite extensive experience our participants had with direct time comparisons in everyday life meant that they were able to make better use of their internal stopwatches in the single-target interval conditions. Perhaps they had the necessary cognitive/neural architecture to succeed in the double and triple conditions, but had not had time to learn to apply it judiciously. In our second experiment we gave our participants additional practice on the double-target interval task, and introduced feedback to make this practice effective. Because this feedback might have led to the development of a strategy for beating the double condition without needing to pause an internal clock (we expand on this later) we no longer split the total target time into equal parts in double-target conditions, instead breaking it up at a random point.

Methods

Methods were identical to Experiment 1 except as outlined below.

Participants

We initially looked at data across a sample of 16 participants[5], having already rejected one participant immediately because no fit could be obtained in one condition (the slope of the fitted function was negative, implying worse than chance performance). In line with Experiment 1, we rejected a further three participants who generated outlying estimates (> 2 SD from mean) for either $\sigma^2_{Increase}$ or *c.*[6] These participants were replaced to produce a final sample of 11 women and five men (mean age = 25.3, SD = 6.8).

Design and Procedure

We used a 2x2 repeated measures design, with only single and double target conditions tested at 300 and 600 ms. On every trial, participants received accurate feedback ("correct", "wrong", or "identical") about their binary judgements. In the single-target condition, there was a 1500 ms silent period between the end of the target stimulus and the start of the standard stimulus. In the double condition, this period was only 1000 ms long (with a 500 ms break between the two target stimuli), equating overall trial duration. Total on time was now divided unequally between the two target stimuli of the double condition: The break could come 30-70% of the way into the combined duration (randomly selected from a uniform distribution on each trial). Importantly, before the counterbalanced presentation of single and double-target stimulus blocks commenced, all participants received a full block of practice at the double-target task.

[5] The slightly larger sample in this experiment reflected our prior expectation that the effect size might be reduced.

[6] All rejected participants had data in line with our main result, i.e. higher values of $\sigma^2_{Increase}$ compared to estimates of c. A wilcoxon test comparing these values in the initial sample was significant ($p = 0.008$).

Data Analysis

Because the triple condition was dropped from this experiment, the value of $\sigma^2_{Increase}$ was calculated by taking the average of two changes in $\sigma^2_{Observed}$: from single to double conditions, at standard durations of both 300 and 600ms.

Results and Discussion

Figure 3 panels A-C show the results of our second experiment. The format is the same as that used for Experiment 1 in Figure 2 (panels C-E). Once again, judgement uncertainty increased from single to double conditions ($F_{[1,15]} = 12.5$, $p = 0.003$) and from short to long intervals (this time significantly; $F_{[1,15]} = 9.4$, $p = 0.008$) with no interaction. Points of subjective equality appeared to diverge from baseline durations and varied across conditions (somewhat surprisingly, given the provision of feedback). Target intervals where subjectively shorter relative to standard intervals in double compared to single conditions ($F_{[1,15]} = 6.9$,$p = 0.019$) and in long compared to short standard conditions ($F_{[1,15]} = 22.4$, $p < 0.001$).These factors did not interact.

<INSERT FIGURE 3 AROUND HERE>

Most critically, the mean value of $\sigma^2_{Increase}$ once again exceeded the mean value of *c* (Figure 3 panel C; $t_{[15]} = 2.2$, $p = 0.041$).[7] This result shows that even with the unfair advantage provided by 120 trials of additional practice on the broken-interval task, participants' performance still deteriorated (relative to a typical comparison task) to a greater extent than stopwatch models would predict.

[7] As noted in footnote 4, we used an approximately unbiased estimator of non-scalar variance. Biasing the estimator by replacing individual negative estimates with zero meant that this result became non-significant at the two-tailed level (although remaining significant if considered as a directional hypothesis).

Experiment 3

Experiments 1 and 2 provided evidence that performance on interval comparison tasks deteriorates sharply when the target is broken up so that the putative internal stopwatch must be paused and restarted. This result makes it unlikely that the model, as currently specified, can fully account for human timing in the sub-second range. However, we felt it important to also consider a recently proposed variant (Taatgen et al., 2007) which embeds a stopwatch timing component within the broad-ranging "adaptive control of thought—rational" (ACT–R) architecture. Although the authors did not conceive their work to be applicable to sub-second intervals (H. Van Rijn, personal communication), the timer they propose might reasonably be adopted by others as a suitable mechanism to explain the pattern of data in our initial experiments. For our purposes, the critical feature of this variant is that in place of a pacemaker with an (on average) constant tick rate, the authors propose a tick rate that slows over time. This proposal has the attractive property of generating scalar variance without the rather arbitrary seeming multiplicative noise that has been incorporated into SET (a feature for which SET has been explicitly criticised; Staddon & Higa, 2006) although it still requires a noise term that is proportional to the inter-tick interval.

If pacemaker rate declines across a timed interval, clear predictions emerge about what should happen to an observer's precision if the interval is split into two or more parts. If we assume that tick rate is reset when the clock is merely paused, the granularity of time will be returned to its initial high acuity state for the second part of the interval, so resolution/precision will actually improve (relative to the typical comparison task without any split), although accuracy will be worse (see below). As this prediction about precision is clearly violated in our first two experiments we did not consider it further here. Similarly, if we assume that click rate is somehow maintained through the break and then begins to decline again, the prediction becomes identical to that of the constant rate

clock that has already been found wanting. However, if we assume that tick rate is not reset unless a new interval estimate is required, and thus that the pacemaker continues to slow during the break in split interval-conditions, we can predict that temporal resolution will be lower in the second part of a split-target interval compared to the equivalent segment of a whole interval. Note that this prediction is qualitatively in line with the results of Experiments 1 and 2. The prediction is schematised in Figure 4.

<INSERT FIGURE 4 AROUND HERE>

Unlike a constant-rate stopwatch, a stopwatch with a decreasing rate also makes strong predictions about accuracy (measured here using the PSE) in split-interval conditions. If clock rate has declined following a break, on average less pulses will be accumulated in the second half of a split interval compared to the same epoch within a contiguous interval. The result would be that the split interval should be perceived as shorter, which, in our experiments, would manifest as an increase in the PSE. This prediction also finds qualitative support in the results of Experiments 1 and 2, at least when considering the change from single (unsplit) to double (split) conditions (see Figures 2 and 3). Does the human brain therefore time short intervals with a stopwatch that can be paused after all, just one with a slowing pacemaker?

To investigate this possibility, we designed an experiment providing additional predictions specific to a slowing pacemaker. In addition to comparing whole and split target stimuli, we now manipulated the duration of the break between the two components of a split-target stimulus. A longer break should exaggerate the changes in precision and accuracy predicted by a slowing stopwatch, because tick rate will have declined even further across a longer gap.

We were additionally concerned that the apparent support for the slowing stopwatch obtained from the PSE data of Experiments 1 and 2 might have reflected something else about the specific context of those experiments. In this paper we have thus far focussed on precision rather than accuracy, in part because accuracy, typically measured as the PSE, is prone to shift dramatically for many reasons in interval timing tasks (e.g. Droit-Volet & Meck, 2007; Heron et al., 2012; Johnston, Arnold, & Nishida, 2006; Penton-Voak, Edwards, Percival, & Wearden, 1996; Tse, Intriligator, Rivest, & Cavanagh, 2004; Wearden, Edwards, Fakhri, & Percival, 1998; Yarrow, Haggard, Heal, Brown, & Rothwell, 2001; Yarrow, Haggard, & Rothwell, 2004; Yarrow, Johnson, Haggard, & Rothwell, 2004; Yarrow & Rothwell, 2003; Zakay & Block, 1996). Indeed, it is almost harder to find an experimental manipulation that does not affect the interval timing PSE than one that does. Of particular concern here, the order of two judged intervals is well known to affect relative judgements about duration (the so called "time order error"; see Hellstroem, 1985, for review) and additional substantial position-dependent biases emerge for short trains of >2 stimuli (e.g. Nakajima, Hoopen, & Van der Wilk, 1991; Rose & Summers, 1995).

These concerns prompted us to verify that PSE shifts were as predicted regardless of these other possible biases. Hence in Experiment 3 we reversed the presentation order of the standard and target intervals so that the standard interval was always presented first. This adaptation also tends to generate a higher degree of precision on interval timing tasks (Dyjas, Bausenhart, & Ulrich, 2012). Finally, in order to confirm that the effects in Experiments 1 and 2 were not specific to a particular type of stimulus, the filled interval used in these initial experiments was substituted with an empty interval. This manipulation should also tend to increase interval timing precision (Grondin, 1993).

Methods

Methods were identical to Experiment 1 except as outlined below.

Participants

A total of 24 participants (two groups of 12; see design) were initially paid to participate. One participant was excluded from the study and replaced due to a negative slope of the fitted function. In line with Experiment 1, we rejected and replaced a further four participants who generated outlying estimates (> 2 SD from mean for their group) for either $\sigma^2_{Increase}$ or *c*.[8] This generated a final sample including 14 women and 10 men (mean age =31, SD = 10.2).

Apparatus and Stimuli

To add generality to our findings, in place of the continuous (filled) intervals used in the first two experiments, 10 ms tones of varying frequencies (500 Hz/1000 Hz/2000 Hz) were now used to clearly demarcate the beginning and end of each empty interval (different frequencies were used to avoid confusion between the intervals to be timed and the gaps between them). In split-target trials, the standard interval was marked by 1000 Hz tones, the first component of the target interval by 500 Hz tones, and the second component of the target interval by 2000 Hz tones. In single-target trials, the 2000 Hz tones were omitted.

Design and procedure

A 2x2x3 design included the between-subject factor *timecourse* and two within-subject factors, *standard duration* (300 & 600 ms) and *target stimulus* (single interval; split interval-short pause [500ms]; split interval-long pause [1500ms]). The presentation order was reversed from the first two

[8] In Experiments 1 and 2 we indicated that the exclusion of participants had no effect by reporting non-parametric tests for the original sample. However, given the more complicated statistical analyses in this experiment (including factorial comparisons) we did not attempt a non-parametric equivalent in this case.

experiments, with the standard duration stimulus now always presented first followed by the target interval(s), thereby exploring the extent to which contextual factors might have generated the PSE changes observed in Experiments 1 and 2. The two separate timecourse groups received slightly different inter-stimulus intervals (see Figure 1c & d). Group A received the standard followed by a 1000 ms gap before the test(s), which meant there were differences in overall trial duration between the baseline, split-short and split-long conditions. Because a potential source of variability in interval judgements can be attributed to memory, with noise potentially accumulating across the period for which an interval estimate must be maintained (e.g. Gamache & Grondin, 2010), in group B the overall trial duration was held constant, achieved by inserting a long (2000 ms) pause between the standard and the first target component when the gap between the two target components was short (500 ms), and a shorter (1000 ms) pause when the gap was long (1500 ms; see Figure 1d). In both groups, the order of the three blocks (single interval, split interval - short pause, split interval - long pause) was counterbalanced across participants. As in Experiment 2, the split-interval durations were randomly subdivided (the first part could be anywhere from 30%-70% of the total with the second part making up the remainder), but no feedback about correctness was provided.

Data Analysis

Our first analysis in this experiment was a repeat of that applied in Experiments 1 and 2, comparing $\sigma^2_{Increase}$ (calculated here by taking the average of four changes in $\sigma^2_{Observed}$: From single to both split-short and split-long conditions, at standard durations of both 300 and 600 ms) to *c*. However, as our primary interest was no longer the SET stopwatch, but rather the slowing-pacemaker stopwatch variant proposed by Taatgen, van Rijn and Anderson (2007), we now additionally simulated this model using Matlab. Taatgen, van Rijn and Anderson provided the following formula describing how the inter-tick interval (t) of the clock evolves over time (pg. 581):

$$t_{n+1} = at_n + noise(M = 0, SD = b \cdot at_n). \quad (7)$$

Here, "noise" indicates a logistic distribution with (M)ean of 0, and the model has 3 free parameters: *startpulse* (t_0, the initial value for the inter-tick interval), *a* (which controls the rate of slowing of the clock) and *b* (which scales the logistic noise). The model was implemented (by simulation) exactly as described by Taatgen, van Rijn and Anderson except in the following regard: Because the slowing pacemaker model was originally validated against data from a task looking at much longer intervals (from ~2-21 seconds, where non-scalar variance makes a negligible contribution) it does not include a term to capture non-scalar variance. Hence for consistency with our overall framework, we adjusted the model slightly by adding a fourth parameter representing Gaussian noise in the (differential) delay to close/open the accumulation switch.

Our approach then followed from the logic applied when testing the SET stopwatch in Experiments 1 and 2, but was elaborated to deal with the slowing-pacemaker model. We first maximum-likelihood fitted the slowing-pacemaker model to the $\sigma_{Observed}$ data from the two single-target conditions (i.e. from a typical interval comparison task, with 300 and 600 ms standards). This fit was obtained individually for each participant, with model predictions generated by simulating an experiment with 21 target levels x 10,000 trials per level x2 standard durations. These simulations yielded trial-by-trial judgements which were treated like real data in our experiments, i.e. converted to probability judged longer at each target level and then fitted with a sigmoid to extract $\sigma_{Observed.}$ Simulations at many different parameter combinations were tested in order to select best-fitting parameters using the Nelder-Mead simplex algorithm (Nelder & Mead, 1965; O'Neill, 1971).

We then used the resulting best-fitting parameters to predict $\sigma_{Observed}$ (and also PSE values) for the split-short and split-long conditions (again, for each participant separately, based on Monte Carlo simulations), assuming that the pacemaker continued to slow during the break. Finally, from these

values, we calculated $\sigma^2_{Increase}$ (and ΔPSE) representing the predicted *change in variance* (and change in PSE) from single to split-short and split-long conditions (we used variance rather than SD changes for consistency with Experiments 1 and 2). Because the slowing-pacemaker model makes different predictions for split-short and split-long conditions, we did not average across these values (as in previous experiments), instead making use of factorial ANOVAs to compare model predictions to the data separately for different split durations and standard intervals.

Results

We first tested the ability of a typical (constant rate) stopwatch to account for the loss of precision in split-interval conditions. Similarly to Experiments 1 and 2, the mean value of $\sigma^2_{Increase}$ significantly exceeded the mean value of *c* (46,522 versus 3,535, $F_{[1,22]} = 10.42$, p=0.004; no effect of group A vs. B or interaction with this factor). Hence the performance deterioration between the whole and the two split-interval conditions cannot be attributed to just the variance associated with switch operations.

Our main concern in Experiment 3, however, was to test the predictions of a *slowing* stopwatch for which there should be a decrease in precision from the short-split to the long-split conditions (as well as from the single-interval to the short-split conditions). As the pulses become more distributed, the point of subjective equality should also change, and should increase, with a greater change emerging for a longer split.

The relevant data are presented in Figure 5 panels a and b. Data have been collapsed across groups A and B as this between-subjects factor was not significant and did not interact with any other factor in any of our ANOVAs. Judgement uncertainty ($\sigma^2_{Observed}$) increased between the short (300ms) and long (600ms) standard intervals ($F_{[1,22]} = 8.63$, p=0.008) and, more importantly, increased from whole

to short-split (500ms gap) to long-split (1500ms) conditions ($F_{[1.1,24.9]}$ = 8.40, p = 0.006; linear trend p = 0.005). Hence the prediction of the slowing stopwatch model was confirmed, at least qualitatively, when considering precision. No other main effects or interactions were significant.

<INSERT FIGURE 5 AROUND HERE>

Points of subjective equality varied between short (300ms) and long (600ms) standard intervals ($F_{[1,22]}$ = 61.90, p <0.001). Note that the direction of this effect was opposite to that obtained in Experiments 1 and 2, with long-standard conditions now inducing a perceived relative *lengthening* of the target stimulus. This suggests that order/context was strongly influencing PSEs. For the critical conditions manipulating the target, the means actually ran in the *opposite* direction to that predicted by the slowing stopwatch model, although this trend was not significant, and there were no interactions with it. In summary, a slowing pacemaker successfully predicts that performance should decline, in terms of precision, from the short-split to the long-split condition. However, as the pulses become more distributed the point of subjective equality should increase, opposite to what we observed.

Qualitative model predictions are useful, but the slowing clock model is sufficiently well specified to permit more precise tests. To provide additional quantitative rigour, we therefore also tested the slowing pacemaker model by finding the best-fitting parameters for each participant in the single-target conditions, and then using these parameters to generate precise predictions for the split-interval conditions (which could be compared to the actual data).

As expected, the model was able to provide a good fit to the single-target conditions (mean ± SD for predicted $\sigma^2_{Observed}$ of 77 ± 40 and 117 ± 67 ms for 300 ms and 600 ms standards respectively, compared to empirical values of 79 ±40 and 120 ± 79 ms). Considering the best fitting parameters

themselves, mean estimates were a close match to those estimated by Taatgen, van Rijn and Anderson using a fit to a very different data set (11.4 ms, 1.19 and 0.014 for *startpulse*, *a* and *b* respectively in our data, compared to 11 ms, 1.1 and 0.015 in their fit, although this close match may have partly reflected our choice to use their estimates to initialise our parameter searches).[9] We also included an additional fourth parameter, reflecting non-scalar variance, and found best-fitting estimates in good agreement with the method applied in Experiments 1 and 2 (where *c* was calculated based on extrapolating scalar variance through a notional duration of zero). Non-scalar variance estimates had a mean of 1925 ms^2 (for the slowing stopwatch parameter fit) vs. 3535 ms^2 (for *c*).

Having obtained fits in single-target (baseline) conditions, we went on to generate model predictions for how much precision should deteriorate ($\sigma^2_{Increase}$) and the PSE should shift (ΔPSE) when a short or long break in the target interval was introduced (i.e. in the short-split and long-split conditions). These data are presented in Figure 5 panels C and D, again shown for all 24 participants because the between-subject factor (timecourse) was not significant and did not interact with any other factors. For precision (see panel C) the general trend is for mean decrements in performance that exceed the predictions of the model. An ANOVA comparing the empirical change in precision with the predictions of the slowing-stopwatch model revealed a significant main effect of model vs. data ($F_{[1, 22]} = 5.03$, $p = 0.032$). Hence we conclude that a stopwatch that can be paused but has a slowing pacemaker does not predict the magnitude of decreased precision that we have observed for sub-second intervals. There was also an interaction between this effect and the duration of the break ($F_{[1, 22]} = 4.40$, $p = 0.048$). The interaction indicates greater violations of model predictions in the long-split ($p = 0.033$) than the short-split ($p = 0.191$) conditions.

[9] Although a longer value of *startpulse* was subsequently used by van Rijn & Taatgen (2008), this would not have been appropriate for our experiments with much shorter intervals. A large value of startpulse generates steps in the psychometric function for short intervals, which are not generally observed, so a model parameterised this way is clearly designed to deal exclusively with longer intervals. We nonetheless felt that the principle of a slowing pacemaker might reasonable be invoked for short-interval timing.

For accuracy (ΔPSE; Figure 5 panel D) the divergence from model predictions is even more striking, with effects in the opposite direction to the model predictions. Here, ANOVA revealed a significant effect of model vs. data ($F_{[1,11]} = 148.13$, $p < 0.001$), with interactions suggesting that this divergence was more pronounced for longer breaks ($F_{[1,11]} = 19.52$, $p < 0.001$) and with longer standards ($F_{[1,11]} = 29.49$, $p < 0.01$). Hence, when considering both precision and accuracy, the data are not supportive of the notion of an internal stopwatch with a slowing pacemaker for intervals at sub-second timescales.

Further simulations: Strategic solutions using a clock that cannot be paused

So far, we have argued that the data from our three experiments are a poor match to the predictions of a pacemaker-accumulator internal stopwatch that can be paused at will. However, models are generally supplanted only when other accounts exist that can provide a better account of existing data. There is of course no shortage of potential models predicting a collapse of performance in split-interval conditions. For example, many models of interval timing do not employ a linear metric, and of particular relevance, many models do not offer any obvious system by which a timing operation could be paused.

To take one prominent example, the striatal beat-frequency model (Matell & Meck, 2004) proposes that a population of neurones in the prefrontal cortex act as oscillators, sending periodic signals to the striatum. Dopiminergic reward signals can generate a memory for a specific interval, equivalent to the coincident pattern of inputs that uniquely specifies that duration, via a mechanism of long-term potentiation. It seems implausible that the oscillatory activity of pre-frontal neurons can be paused and then resumed from the saved state at will. Furthermore, there is no continuous metric

for time in this model and thus no obvious way to perform temporal arithmetic: The coincident patterns that would specify each of two short intervals would have no consistent translation into the coincident pattern that would specify their sum. Hence the model predicts that performance should plummet in broken-interval conditions. A similar analysis could be applied to models based on exponentially-decaying memory traces (Staddon & Higa, 1999) or neural network dynamics (Buonomano & Merzenich, 1995).

At first glance, however, our data appear just as problematic for these models as they do for stopwatch models, because while performance certainly got a lot worse in split-interval conditions, most participants could still perform well above chance. Yet with a little thought, it is quite straightforward to come up with some plausible strategies by which an observer might achieve this level of performance without having to pause an internal clock. Indeed, some aspects of our data seem highly suggestive of strategic solutions. For example, while some degree of between-subject variability in the ability to time intervals is to be expected, it is noticeable that this variability is greatly magnified in split-interval conditions compared to the single-target baseline condition (see, for example, the error bars in Figure 1c). Clearly some participants were very much better able to handle split-interval tasks than others, to an extent that greatly outstripped differences in basic timing ability. This seems more consistent with differences in the ability to find a workable strategy than with all participants making the best use they can of the same hard-wired clock. We outline two possible strategies next, and then simulate one in order to demonstrate that it could potentially account for some of our findings.

In Experiment 1, participants might have ignored the standard stimulus altogether and attended to just one segment of the target stimulus (rather than attempting to sum all segments together). In this way they would convert the task into a psychophysical procedure known as the method of single stimuli. By comparing a single sub-interval to an average duration for this same sub-interval built up

across trials and maintained in long-term memory, competent performance might be achieved (Morgan et al., 2000). This strategy would probably have been weakened in Experiment 1 by our interleaving of two different standards, and further undermined in Experiment 2 onwards, where broken-interval stimuli had randomised durations, thus introducing external noise. We do not consider it likely that the majority of participants stumbled upon this strategy, because it actually has the potential to yield a performance *improvement* relative to baseline, at least in Experiment 1 (by allowing participants to effectively time shorter durations) and doesn't offer any account of the larger decrements found with a longer break in the split-interval conditions of Experiment 3.

As a second possibility, the observer might, like our first proposal, ignore the standard in order to treat the task as effectively a method of single stimulus presentation. However, instead of attempting to compare just one component of the split interval to its own average across trials, participants might time the *entire* interval from onset of the first part of the split target to offset of its final component. Because we did not randomise the break duration in any of our experiments, this interval gives unambiguous information about this trial's target duration relative to previous trials' target durations, and splitting the target unequally does nothing to defeat it. However, this strategy forces the observer to time much longer intervals, which are subject to decrements in precision as per Weber's law. Unlike the previous strategy, it predicts worse performance with a longer break in split-interval conditions, as obtained in Experiment 3, and with a triple rather than double stimulus, as obtained in Experiment 1.

We have outlined two strategies, but many others can easily be conceived (for example, timing just one target sub-stimulus in the broken-interval condition and then attempting to run the clock two or three times in rapid succession during the subsequent standard interval, or performing a comparison between the longer of the split target's components and the standard, and responding "longer" if

either sub-stimulus on its own even *approaches* parity with the standard). It is thus evident that the space of possible strategic solutions to the split-interval task is very large, particularly when you consider that a given set of participants are probably mixing and matching different approaches, and that modelling any strategy implies specifying it in considerable detail such that parameterisation (or at least some judicious decisions during implementation) would tend to give even greater scope for success. It is also somewhat unfair to test stopwatch models against a new situation (the sub-second split-interval task) for which they were (arguably) not designed, and for which they have not been elaborated, and then conclude that strategic alternatives are better simply because they might explain this new situation in an entirely post-hoc manner. However, it does seem worthwhile to at least demonstrate the plausibility of these sorts of strategies to approximate the kind of data patterns we observed without recourse to an internal stopwatch that can be paused. To this end, we simulated the second of these strategic solutions making use of a generic clock that cannot be paused, as follows.

Methods

We worked with the average estimates obtained from Experiment 3.[10] Mean non-scalar variance and scalar increases in variance (obtained by fitting 300 ms and 600 ms baseline performance to a generalised version of Weber's law) were used to predict performance at longer intervals (800, 1100, 1800 and 2100 ms) corresponding to the periods incorporating the target stimulus *and the breaks* in the short-break and long-break conditions. We arbitrarily assumed that participants could

[10] The participant-by-participant approach was less viable for this simulation, because noise at the individual participant level yields a subset of participants who performed slightly better in 600 ms than 300 ms baseline conditions, such that extrapolation via Weber's law to very long interval timing predicts substantially sub-zero variance.

build an internal standard based on the average of all previously experienced trials in a block.[11] On this basis, we partitioned the variance extrapolated from *baseline* conditions into a large component contributed by the target stimulus and a much smaller component contributed by the average of all standard stimuli.[12] We then performed 1000 Monte-Carlo simulations of our exact experimental procedure (60 trials at each standard duration, selected from an adaptive distribution on each trial) to estimate precision in *split-interval* conditions, where we assumed that participants compared the entire interval (from target component #1 onset to target component #2 offset) to an internal standard formed by averaging this same interval across all previous trials.[13]

<INSERT FIGURE 6 AROUND HERE>

Results of these simulations are shown in Figure 6, alongside the predictions of the slowing-stopwatch model and the data obtained in Experiment 3. Did observers use a strategy of timing both target-stimulus components and also the break between them? Clearly not in all cases, as mean performance did not decline as much as this strategy predicts, particularly for the 300 ms standard and long-split condition. Note, however that there is some evidence suggesting that by extrapolating from sub-second stimulus durations to durations around 2000 ms using an assumption of scalar variance we could have overestimated judgement variability (Lewis & Miall, 2009; but see Bangert et al., 2011). Less controversially, if participants adopted a counting strategy, which is likely at supra-second intervals, we would certainly have overestimated the likely drop in performance (Getty,

[11] Evidence from method of single stimulus experiments suggests that this is not unreasonable (Morgan, Watamaniuk & McKee, 2000) although those authors were not looking at timing, and used experienced observers as participants.

[12] Essentially, we assumed that in whole-interval (baseline) conditions, participants formed an average of the first interval, and compared it to the estimate obtained on each trial for the second interval. There is evidence to suggest that observers use a strategy similar to this in interval comparison (Dyjas, Bausenhart & Ulrich, 2012) although those authors used a slightly different scheme for creating the internal standard (a geometric moving average).

[13] We used the median of our 1000 simulated values for precision rather than the mean, having first excluded simulations returning a negative slope, as a few very extreme values can strongly affect the mean. This choice is justified by our experimental procedures, as we also excluded participants for whom slopes were negative, or who returned extreme outlying values.

1976; Grondin, Meilleur-Wells, & Lachance, 1999; Grondin, Ouellet, & Roussel, 2004). One could therefore obtain something approaching our data if many participants use a strategy such as timing the whole period incorporating all segments of the target stimulus and the break, particularly if a second subset of participants used one of the better strategies that were also available.[14]

General Discussion

In our experiments, we first measured participant performance in a classic interval comparison task. We then employed the scalar property to estimate the maximum decrement in precision which should arise as a result of the additional mental operations required under a stopwatch model in our novel split-interval tasks. The results revealed a decline in performance which greatly exceeded the calculated estimate in all three experiments, suggesting that participants cannot pause and restart the accumulation of temporal pulses like a stopwatch when timing short intervals. This was true even with the intensive training and feedback provided in Experiment 2.

Next, we looked at a more recent stopwatch architecture, focusing on a crucial feature of this model - the gradually slowing pacemaker. Although lengthening the gaps in our split-interval task caused participants' performance to decline, as would be expected under this model, the target interval was perceived as longer than an unbroken standard, rather than shorter, which is in direct contrast to the results predicted by a slowing pacemaker. Clearly, patterns obtained in Experiments 1 and 2 (where the standard terminated the trial) were much more compatible with a slowing-pacemaker account than those obtained in Experiment 3 (where the standard initiated the trial). Hence a fairer test might have been to attempt to remove the contextual bias on PSE before making any

[14] For example, an identical Monte-Carlo method used to predict precision using the *first* strategy we outlined (comparing the first sub-component of the split standard to the average such stimuli over previous trials) yielded $\sigma^2_{Increase}$ values of ~14,000 and ~23,000 for 300 and 600 ms standards respectively, regardless of break duration.

comparison with model predictions, perhaps by running conditions with standards in both initial and terminal positions and taking the average PSE shifts.

However, even if we take into account the close association which exists between subjective duration and the context in which an interval is presented (e.g. Rose & Summers, 1995) and thus ignore the data regarding accuracy, the magnitude of the increase in judgement uncertainty (particularly prominent in the long-split condition) deviated significantly from model predictions in Experiments 3, suggesting a slowing pacemaker account is not sufficient to explain our data. It has previously been shown that the ACT-R module with slowing pacemaker is able to predict behaviour in a variety of supra-second timing tasks (van Rijn & Taatgen, 2008) ranging from simple discrimination to the timing of multiple intervals, via a single timing mechanism. The results of our final experiment question this kind of model's applicability specifically for sub-second intervals.

Although performance was poor compared to stopwatch predictions, participants were nonetheless able to complete our split-interval tasks with varying rates of success, presumably denoting the use of strategies capable of achieving moderate performance levels alongside an internal clock that cannot be paused (e.g. Buonomano & Merzenich, 1995; Matell & Meck, 2004). We looked for the best fit for the single most feasible participant strategy (i.e. timing the entire interval, including the break). While we did not attempt a formal statistical comparison in this case, it was clear that this strategy underperforms when set against the group's precision, so we speculate that certain participants employed superior strategies to this one. This analysis is useful, however, because it demonstrates that participants can in principle perform the split-interval task even without an internal clock that can be paused. We feel that a careful consideration of the various strategies available in timing tasks is critical when assessing the meaning of psychophysical data. Regrettably, we did not ask systematically about the strategies our participants were using (although

introspection may not have been very revealing in any case). We would expect a range of strategies, based on individual differences and problem solving capabilities.

Although we suspect strongly that our participants performed strategically, using a timer that cannot be paused, it is worth emphasising that in our strategic solutions we considered only precision (the main focus of our paper) and not accuracy. Stopwatch models do not explain our patterns of PSEs, which were clearly subject to strong contextual biases, but neither do our proposed strategies. We acknowledge that this represents a weakness of our presentation (in rejecting one model, one should ideally specify a model which *fully* explains the data in its place). However, the range and complexity of biasing effects in interval timing will clearly require substantial add-ons to any basic clock model.

Following from this point, it is certainly possible to construct plausible accounts of some of our observed biases. For example, there is evidence suggesting that participants typically form an averaged internal standard, preferentially for the *first* comparison stimulus (Dyjas et al., 2012). When considered alongside our mixing of two different standards within a single block, this process can provide an explanation for the pattern of biases observed in baseline (single-target) conditions (i.e. relative temporal dilation of the target at one standard duration and contraction for the other, with these biases flipping when the standard changed position from Experiments 1 & 2 to Experiment 3). The additional biases evident in split-interval conditions could perhaps be explained with reference to alternative forms of assimilation/contrast, although the way the addition of a break affected PSEs oppositely in Experiments 1 & 2 versus Experiment 3 would require a fairly inventive formulation. We will not pursue such accounts further, as we are dubious regarding whether any attempt to explain all the biases we observed here would generalise beyond our current data set.

Relationship to previous split-interval literature

This is certainly not the first study to make use of broken intervals to investigate the mechanisms of interval timing, and our conclusions may seem at odds with previous work where humans have successfully interpreted broken intervals (e.g . Fortin & Tremblay, 2006; Fortin et al., 2009; Tremblay & Fortin, 2003). However, there are two important features of these experiments that may resolve the apparent conflict with our findings. Firstly, the aforementioned studies have used supra-second intervals. These would permit counting or sub-dividing strategies (Grondin et al., 1999; Grondin et al., 2004), a key reason we chose to investigate shorter durations in this paper. Clearly, if one can count, one can ignore a gap even with a clock that cannot be paused. Related to this issue, previous gap studies have not discussed strategic solutions to their tasks in detail, or attempted to model them.

Secondly, the focus of previous studies has been on accuracy, not precision (and on the effects of diverted attention on timing, rather than testing the stopwatch model per se). With our detailed consideration of both precision and accuracy, our analyses have a greater capacity to detect deviations from model predictions. We therefore suggest that the ability to pause an internal clock, as inferred in previous work, may be more apparent than real. However, we also recognise that a different kind of clock may be used for longer durations (a point consider further below).

The seeming ability to take account of breaks has also been suggested by non-human animal work. Early studies (e.g. Roberts & Church, 1978) utilised a peak-interval procedure wherein animals receive reinforcement at the to-be-timed duration. Such studies have generally analysed the average response rate on “peak” (non-reinforced) trials, so that responses beyond the point of reinforcement can be assessed. When the stimulus is broken, it sometimes appears as though the animal subjects retain the pre-gap interval in memory and then resume timing where they left off,

consistent with how we would expect a stopwatch model to function. However, such accurate performance is far from a certainty, and rats and pigeons also often behave as though their clocks continue to run straight through the gap, or indeed start again from scratch after it ends. For example, when the PI procedure was reversed (Buhusi & Meck, 2000) and the animals were required to time the absence of a signal, with the gap denoted with a stimulus, the results suggested that the entire timing process was restarted following the gap. Indeed, recent formulations explaining the various patterns of results that animals may exhibit have abandoned the notion of a pause in accumulation (e.g. Buhusi, 2012). Instead, the clock continues to accumulate, but a second process (memory leakage) is posited, the strength of which depends on a variety of stimulus factors. Such a modification of the standard model is certainly interesting[15], but is beyond our scope here, particularly given the much longer (tens of seconds) intervals typically investigated in animal timing.

Following on from this, there are clearly some possible adjustments of human stopwatch accounts that would make them compatible with our data. One might argue that opening the switch and thus stopping accumulation leads to a *mandatory* reset of the accumulator to zero, i.e. the timer is a counter, but one that cannot be paused. This would of course generate some problems for proponents of counter models, for example regarding how they deal with the effects of selective attention on time perception via a "flickering switch" (Lejeune, 1998). Although the capacity to pause timing is not the central defining feature of counter models, it is certainly a feature that provides additional explanatory power, and which has been adopted as an asset in the evidential competition with other proposed pacemakers. Alternatively, the ability to pause the clock could be retained and the models reformulated in other ways, for example by postulating some currently unspecified process of noise accumulation during the pause. This, however, is a substantial departure from basic stopwatch models, and at the very least this would reduce their parsimony.

[15] It might even offer an account for the biasing effects of gaps seen here (i.e. perceptual compression in Experiments 1 & 2 versus expansion in experiment 3), albeit an extremely ad hoc one, in which a gap between filled intervals is for some reason more salient than a gap between empty intervals.

Another reasonable response to our data would be to suggest that stopwatch mechanisms might only operate at longer interval durations than those tested here. Several groups have posited some division of temporal mechanisms, with the dividing point generally placed above the interval durations we tested. For example, Lewis and Miall (2003) have proposed such a distinction on the basis of patterns of brain activations from neuroimaging studies, while Rammsayer has made a similar case based on the differential effect of pharmacological interventions on short and long interval timing (e.g. Rammsayer, 1999). Indeed, many stopwatch accounts appear to have been designed specifically to explain supra-second timing. However, we would emphasise that many theorists have previously applied the stopwatch model at sub-second durations (e.g. Wearden et al., 1998) so even if all we have shown is that the stopwatch does not operate in this particular range, our findings are clearly not trivial.

Before concluding, it is only fair to note that the logic with which we discredit a counter-based stopwatch model in our first two experiments depends on the reality of the generalised version of Weber's law for time perception (see Wearden & Lejeune, 2008, for a recent review of this issue). However, we suspect that alternative means of estimating the non-scalar variance associated with initiating and ending a period of timing would have yielded the same answer. Indeed, although we did not present the analysis here, the performance decrement in our first experiment actually significantly exceeded the *total* (scalar and non-scalar) variability recorded with an unbroken 300 ms standard. Clearly no means of estimating the noise accrued via switch operations could have come up with an estimate that exceeded the total variability associated with the task.

Conclusions

We have investigated participants' ability to perform temporal arithmetic with short sub-second intervals, focussing for the first time on predictions about precision rather than accuracy. If humans can in fact start and stop accumulation from an internal timing process at will, and maintain a time count so as to concatenate brief periods of time divided by a break, then doing so makes them considerably less precise. This is not what stopwatch models (as currently formulated) predict, which implies that they need to be developed/complicated with additional arbitrary seeming processes or constrained to account only for longer intervals.

References

Acerbi, L., Wolpert, D. M., & Vijayakumar, S. (2012). Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Computational Biology*, 8, e1002771.

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

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036.

Bangert, A. S., Reuter-Lorenz, P. A., & Seidler, R. D. (2011). Dissecting the clock: Understanding the mechanisms of timing across tasks and temporal intervals. *Acta Psychologica, 136*(1), 20-34.

Buhusi, C. V. (2012). Time-sharing in rats: Effect of distracter intensity and discriminability. *Journal of Experimental Psychology: Animal Behavior Processes, 38*(1), 30.

Buhusi, C. V., & Meck, W. H. (2000). Timing for the absence of a stimulus: The gap paradigm reversed. *Journal of Experimental Psychology: Animal Behavior Processes, 26*(3), 305.

Buhusi, C. V., & Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews Neuroscience, 6*(10), 755-765.

Buhusi, C. V., & Meck, W. H. (2006). Interval timing with gaps and distracters: Evaluation of the ambiguity, switch, and time-sharing hypotheses. *Journal of Experimental Psychology. Animal Behavior Processes, 32*(3), 329-338.

Buhusi, C. V., Sasaki, A., & Meck, W. H. (2002). Temporal integration as a function of signal and gap intensity in rats (rattus norvegicus) and pigeons (columba livia). *Journal of Comparative Psychology, 116*(4), 381-390.

Buonomano, D. V., & Merzenich, M. M. (1995). Temporal information transformed into a spatial code by a neural network with realistic properties. *Science, 267*(5200), 1028-1030.

Creelman, C. D. (1962). Human discrimination of auditory durations. *Journal of the Acoustical Society of America, 34*, 582-593.

Crystal, J. D. (1999). Systematic nonlinearities in the perception of temporal intervals. *Journal of Experimental Psychology: Animal Behavior Processes, 25*(1), 3-17.

Droit-Volet, S., & Meck, W. H. (2007). How emotions colour our perception of time. *Trends in Cognitive Sciences, 11*(12), 504-513.

Dyjas, O., Bausenhart, K. M., & Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. *Attention, Perception, & Psychophysics, 74*(8), 1819-1841.

Fortin, C., Fairhurst, S., Malapani, C., Morin, C., Towey, J., & Meck, W. H. (2009). Expectancy in humans in multisecond peak-interval timing with gaps. *Attention, Perception, &Psychophysics, 71*(4), 789-802.

Fortin, C., & Masse, N. (2000). Expecting a break in time estimation: Attentional time-sharing without concurrent processing. *Journal of Experimental Psychology: Human Perception and Performance, 26*(6), 1788-1796.

Fortin, C., & Tremblay, S. (2006). Interrupting timing in interval production and discrimination: Similarities and differences. *Behavioural Processes, 71*(2-3), 336-343.

Gamache, P., & Grondin, S. (2010). Sensory-specific clock components and memory mechanisms: Investigation with parallel timing. *European Journal of Neuroscience, 31*(10), 1908-1914.

Getty, D. J. (1976). Counting processes in human timing. *Perception & Psychophysics, 20*(3), 191-197.

Getty, D. J. (1975). Discrimination of short temporal intervals: A comparison of two models. *Perception and Psychophysics, 18*(1), 1-8.

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

Grondin, S. (1993). Duration discrimination of empty and filled intervals marked by auditory and visual signals. *Perception & Psychophysics, 54*(3), 383-394.

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., Meilleur-Wells, G., & Lachance, R. (1999). When to start explicit counting in a time-intervals discrimination task: A critical point in the timing process of humans. *Journal of Experimental Psychology: Human Perception and Performance, 25*(4), 993-1004.

Grondin, S., Ouellet, B., & Roussel, M. (2004). Benefits and limits of explicit counting for discriminating temporal intervals. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale, 58*(1), 1-12.

Grondin, S. (2001). From physical time to the first and second moments of psychological time. *Psychological Bulletin, 127*(1), 22-44.

Hellstroem, A. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin, 97*(1), 35-61.

Heron, J., Aaen-Stockdale, C., Hotchkiss, J., Roach, N. W., McGraw, P. V., & Whitaker, D. (2012). Duration channels mediate human time perception. *Proceedings of the Royal Society Series B: Biological Sciences, 279*(1729), 690-698. doi:10.1098/rspb.2011.1131

Ivry, R. B., & Hazeltine, R. E. (1995). Perception and production of temporal intervals across a range of durations: Evidence for a common timing mechanism. *Journal of Experimental Psychology: Human Perception and Performance, 21*(1), 3-18.

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

Johnston, A., Arnold, D. H., & Nishida, S. (2006). Spatially localized distortions of event time. *Current Biology, 16*(5), 472-479.

Lejeune, H. (1998). Switching or gating? the attentional challenge in cognitive models of psychological time. *Behavioural.Processes, 44*(2), 127-145.

Lewis, P. A., & Miall, R. C. (2003). Distinct systems for automatic and cognitively controlled time measurement: Evidence from neuroimaging. *Current Opinion in Neurobiology, 13*(2), 250-255.

Lewis, P. A., & Miall, R. C. (2009). The precision of temporal judgement: Milliseconds, many minutes, and beyond. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences, 364*(1525), 1897-1905.

Malapani, C., & Fairhurst, S. (2002). Scalar timing in animals and humans. *Learning and Motivation., 33*(1), 156-176.

Matell, M. S., & Meck, W. H. (2004). Cortico-striatal circuits and interval timing: Coincidence detection of oscillatory processes. *Brain Research. Cognitive Brain Research, 21*(2), 139-170.

Matthews, W. J., & Grondin, S. (2012). On the replication of Kristofferson's (1980) quantal timing for duration discrimination: Some learning but no quanta and not much of a weber constant. *Attention, Perception, & Psychophysics, 74*(5), 1056-1072.

Morgan, M. J., Watamaniuk, S. N., & McKee, S. P. (2000). The use of an implicit standard for measuring discrimination thresholds. *Vision Research, 40*(17), 2341-2349.

Nakajima, Y., Hoopen, G. T., & Van der Wilk, R. (1991). A new illusion of time perception. *Music Perception, 8*(4) , 431-448.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308-313.

O'Neill, R. (1971). Algorithm AS 47: Function minimization using a simplex procedure. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 20*(3), 338-345.

Penton-Voak, I. S., Edwards, H., Percival, A., & Wearden, J. H. (1996). Speeding up an internal clock in humans? effects of click trains on subjective duration. *Journal of Experimental Psychology. Animal Behavior Processes, 22*(3), 307-320.

Rammsayer, T. H. (1999). Neuropharmacological evidence for different timing mechanisms in humans. *The Quarterly Journal of Experimental Psychology: Section B, 52*(3), 273-286.

Rammsayer, T.H. & Ulrich, R. (2005). No evidence for qualitative differences in the processing of short and long temporal intervals. *Acta Psychologica, 120*(2), 141-171.

Roberts, S., & Church, R. M. (1978). Control of an internal clock. *Journal of Experimental Psychology: Animal Behavior Processes, 4*(4), 318.

Rose, D., & Summers, J. (1995). Duration illusions in a train of visual stimuli. *Perception, 24*(10), 1177-1187.

Rosenberger, W. F., & Grill, S. E. (1997). A sequential design for psychophysical experiments: An application to estimating timing of sensory events. *Statistics in Medicine, 16*(19), 2245-2260.

Staddon, J. E., & Higa, J. J. (1999). Time and memory: Towards a pacemaker-free theory of interval timing. *Journal of the Experimental Analysis of Behavior, 71*(2), 215-251.

Staddon, J. E., & Higa, J. J. (2006). Interval timing. *Nature Reviews Neuroscience, 7*(8), c1-c2.

Taatgen, N. A., van Rijn H., & Anderson, J. (2007). An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning. *Psychological Review, 114*(3), 577-598.

Taatgen, N., & van Rijn, H. (2011). Traces of times past: representations of temporal intervals in memory. *Memory & Cognition*, 39, 1546-1560.

Treisman, M. (1963). Temporal discrimination and the indifference interval: Implications for a model of the "internal clock." *Psychological Monographs, 77*(13)

Tremblay, S., & Fortin, C. (2003). Break expectancy in duration discrimination. *Journal of Experimental Psychology: Human Perception and Performance, 29*(4), 823-831.

Tse, P. U., Intriligator, J., Rivest, J., & Cavanagh, P. (2004). Attention and the subjective expansion of time. *Perception and Psychophysics, 66*(7), 1171-1189.

van Rijn, H., & Taatgen, N. A. (2008). Timing of multiple overlapping intervals: How many clocks do we have? *Acta Psychologica, 129*(3), 365-375.

Wearden, J. H., Edwards, H., Fakhri, M., & Percival, A. (1998). Why "sounds are judged longer than lights": Application of a model of the internal clock in humans. *Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology, 51*(2), 97-120.

Wearden, J. H., & Lejeune, H. (2008). Scalar properties in human timing: Conformity and violations. *Quarterly Journal of Experimental Psychology, 61*(4), 569-587.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception and Psychophysics., 63*(8), 1293-1313.

Yarrow, K., Haggard, P., Heal, R., Brown, P., & Rothwell, J. C. (2001). Illusory perceptions of space and time preserve cross-saccadic perceptual continuity. *Nature, 414*(6861), 302-305. doi:10.1038/35104551

Yarrow, K., Haggard, P., & Rothwell, J. C. (2004). Action, arousal, and subjective time. *Consciousness and Cognition, 13*(2), 373-390. doi:10.1016/j.concog.2003.10.006

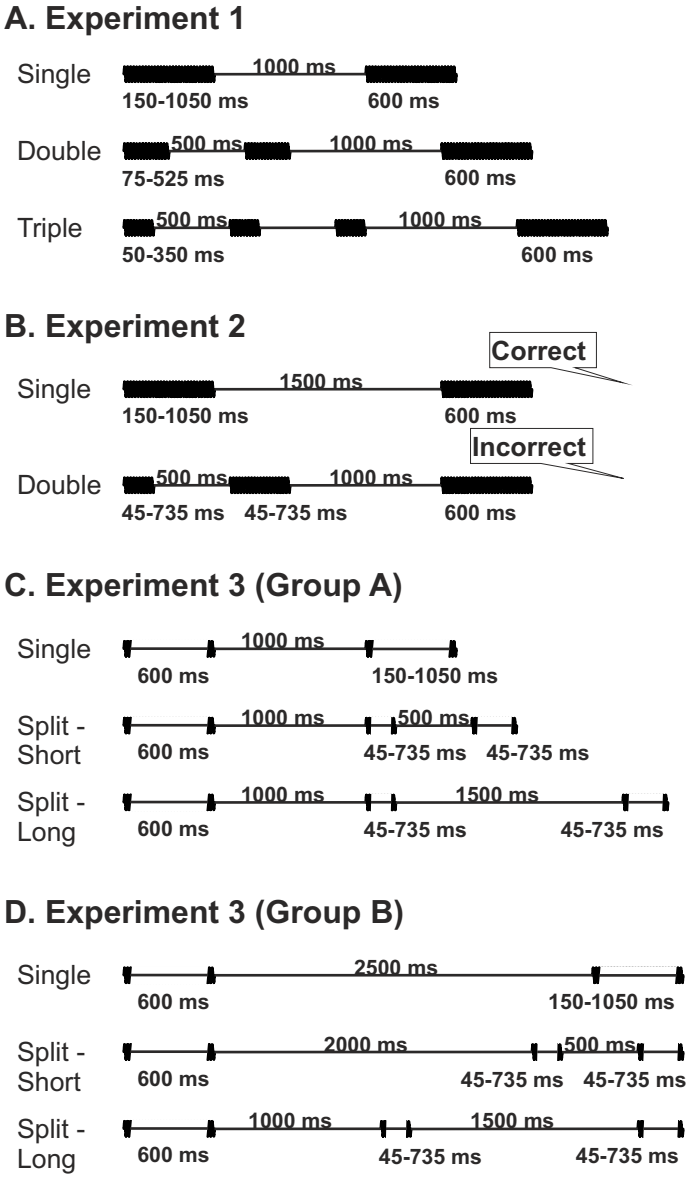
Yarrow, K., Johnson, H., Haggard, P., & Rothwell, J. C. (2004). Consistent chronostasis effects across saccade categories imply a subcortical efferent trigger. *Journal of Cognitive Neuroscience, 16*(5), 839-847. doi:10.1162/089892904970780

Yarrow, K., & Rothwell, J. C. E. (2003). Manual chronostasis: Tactile perception precedes physical contact. *Current Biology, 13*, 1134-1139.

Zakay, D., & Block, R. A. (1996). The role of attention in time estimation processes. *Advances in Psychology, 115*, 143-164.

Zakay, D. (2000). Gating or switching? gating is a better model of prospective timing (a response to 'switching or gating?' by lejeune). *Behavioural Processes, 52*(2-3), 63-69.

Figure 1

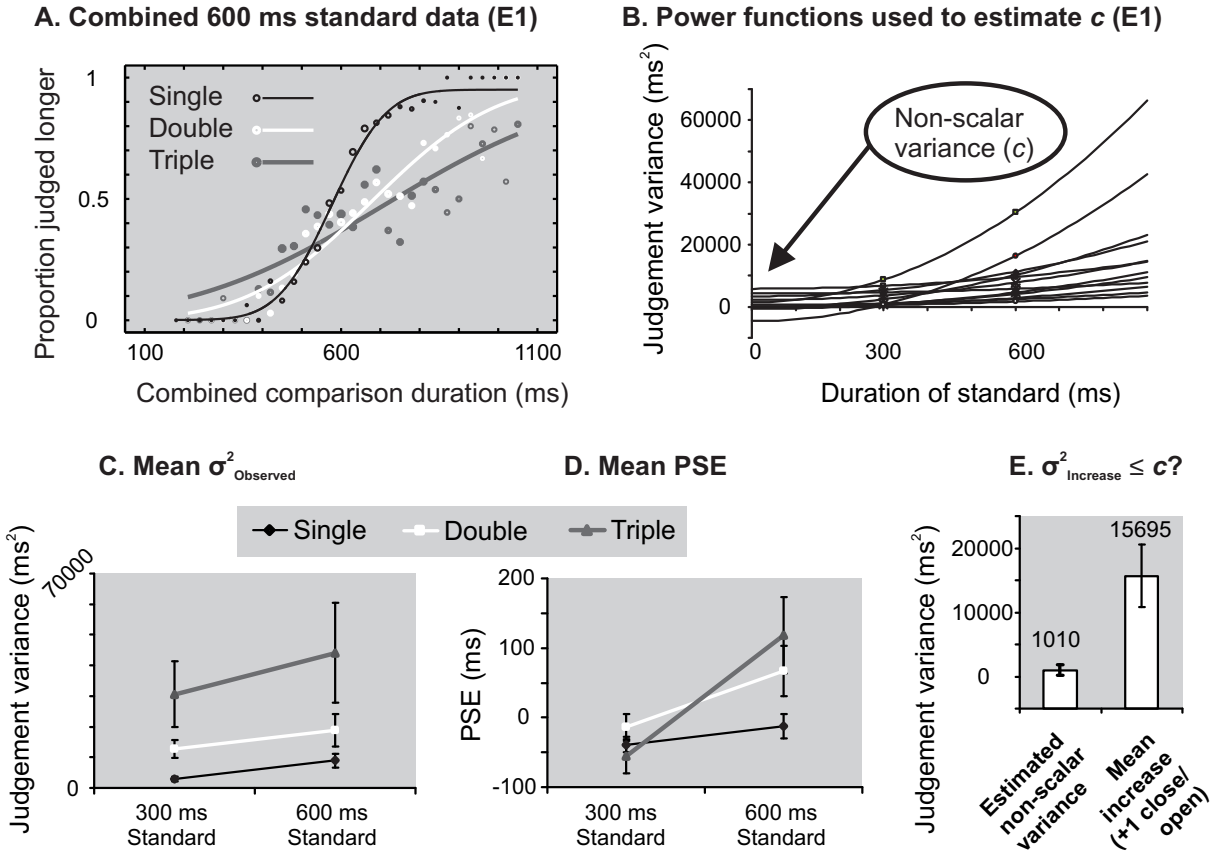


Legend to Figure 1

Schematic timeline of experimental tasks. A. Illustration of the 600 ms standard conditions from Experiment 1. The target segment could be presented whole or divided into two or three separate parts (in different blocks of trials). B. Illustration of the 600 ms standard conditions from Experiment 2. Single and double target stimulus conditions were presented, with trial-by-trial feedback and matched overall trial durations. The two target stimuli in double conditions were of unequal duration, with the total target time divided at random. C. Illustration of the 600 ms standard

conditions from Experiment 3 group A. The standard was presented first, in an empty interval task, and there were two variants of the double condition (split short and split long) which varied the duration of the gap between segments. D. Illustration of the 600 ms standard conditions from Experiment 3 group B. The time from the standard to the onset of the target was co-varied with gap duration in order to equate overall trial duration.

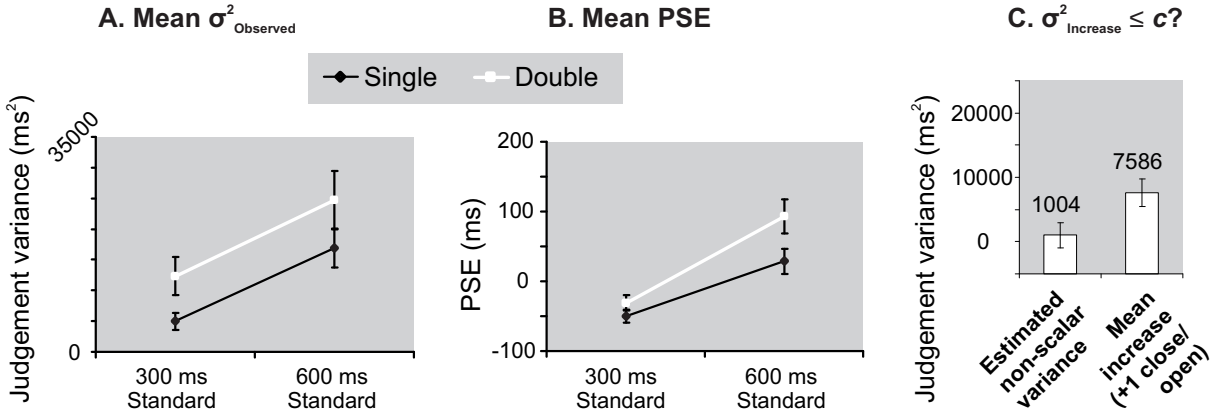
Figure 2



Legend to Figure 2

Results of Experiments 1. **A.** Raw data combined across all participants for the 600 ms standard conditions of Experiment 1. Continuous lines show cumulative Gaussian fits to this data. The size of each data point provides a rough guide to the number of trials recorded for that target duration. Note that the adaptive procedure we employed to select target durations meant that the best performing participants were sampled mainly in the central portion of the figure, so noise at more extreme positions reflects judgements of less able performers. **B.** Judgement variance plotted separately for all participants in the single-target (baseline) conditions of Experiment 1. Power functions are drawn through these data to reveal estimates of non-scalar variance. **C-E.** Mean judgement variance (C), points of subjective equality (shown as deviations from accurate performance, i.e. objective equality would be at zero; D) and comparisons of the mean increase in judgement uncertainty (per break) with *c*, the estimated non-scalar variance (E). Error bars show standard error of the mean.

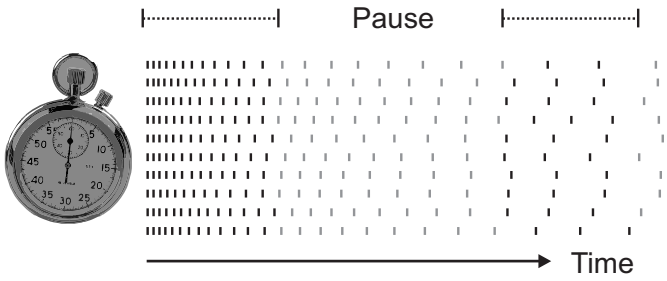
Figure 3



Legend to Figure 3

Results of Experiment 2. **A-C.** Mean judgement variance (A), points of subjective equality (shown as bias relative to objective equality = 0; B) and comparisons of the mean increase in judgement uncertainty (per break) with the estimated non-scalar variance (C). Error bars show standard error of the mean.

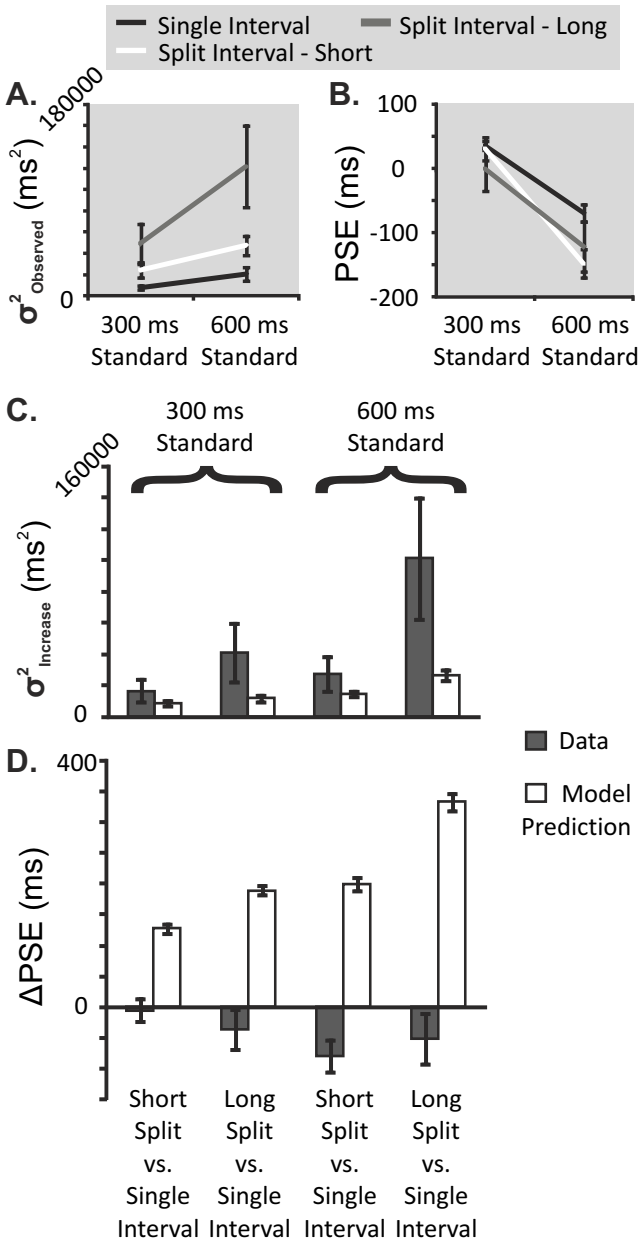
Figure 4



Legend to Figure 4

Schematic illustration of an internal stopwatch with a slowing pacemaker (Taatgen, van Rijn & Anderson, 2007). Timing of a split interval is shown (300/500/300 ms for the first segment/break/second segment, respectively) above ten simulated trials. The ticks that would be accumulated are shown in black and those that would fall in the gap are shown in grey. Notice how the second half of the split has a reduced resolution, capturing less ticks (which has implications for the PSE) and implying that a greater change in interval duration would be required in order to generate a tick count that could be discriminated (which has implications for judgement uncertainty).

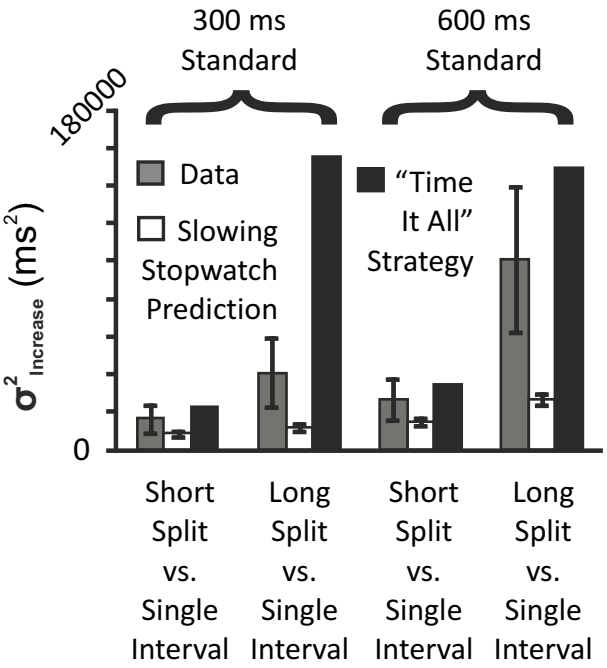
Figure 5



Legend to Figure 5

Results of Experiment 3 averaged across all 24 participants. **A-B.** Mean judgement variance (A) and points of subjective equality (shown as bias relative to objective equality = 0; B). **C-D.** Comparisons of the increases in judgement variance (C) and changes in PSE (D) with those predicted by an internal stopwatch with a slowing pacemaker (where tick rate continues to decline through the pause). Error bars show standard error of the mean.

Figure 6



Legend to Figure 6

Data from Experiments 3 alongside strategic predictions. Empirical increase in judgement variance (grey) is shown alongside the predictions of an internal stopwatch with a slowing pacemaker (white; tick rate continues to decline through the pause) and a second set of predictions based on a strategic solution to the task (black; the observer estimates the entire period of the split target stimulus, including the split, and compares it to an internal standard). Error bars show standard error of the mean (where available).