Modelling Feedback Effects on the Production of Short Time Intervals

John H. Wearden[1,2,*] and Jordan Wehrman[3,,***]**

[1]School of Psychology, Keele University, Keele, ST5 5BG, UK
[2]Division of Neuroscience and Experimental Psychology, Manchester University, Manchester, M13 9PL, UK
[3]Department of Cognitive Science, Macquarie University, 2122 Sydney, Australia

Abstract

People produced time intervals of 500 to 1250 ms, with accurate feedback in ms provided after each production. The mean times produced tracked the target times closely, and the coefficient of variation (standard deviation/mean) declined with increasing target time. The mean absolute change from one trial to another, and its standard deviation, measures of trial-by-trial change, also increased with target time. A model of feedback was fitted to all four measures. It assumed that the time produced resulted from a combination of a scalar timing process and a non-timing process. Although the non-timing process was on average invariant with target time, the timing process was assumed to be sensitive to feedback, in two different ways. If the previous production was close to the target the model repeated it (a repeat process), but if it was further away the next production was adjusted by an amount related to the discrepancy between the previous production and the target (an adjust process). The balance between the two was governed by a threshold, which was on average constant, and it was further assumed that the relative variability of the repeat process was lower than that of the adjust process. The model produced output which fitted three of the four measures well (average deviation of 3 or 4%) but fitted the standard deviation of change less well. Reducing the magnitude of the non-timing process produced output which conformed approximately to scalar timing, and the model could also mimic data resulting from the provision of inaccurate feedback.

[*] ORCID: 0000-0001-6589-4613

[**] To whom correspondence should be addressed. E-mail: jordan.wehrman@mq.edu.au

[***] ORCID: 0000-0002-9358-6092

1. Introduction

The method of interval production, where people are required to produce a specified time interval by making some response, such as holding down a button or pressing a key twice to start and end the interval, is one of the classic trio of timing tasks, along with reproduction and verbal estimation. The present article reports a very simple experiment on the production of time intervals ranging from 500 to 1250 ms, but its main focus is on the effects of feedback, and in particular an attempt to model how feedback controls interval production in the absence of chronometric counting.

Quantitative models of performance on timing tasks such as temporal bisection (Wearden, 1991) or temporal generalization (Wearden, 1992) are now commonplace but, so far as we know, no model of feedback effects on simple interval production without counting has been developed until now. We begin with a brief review of effects of feedback on production tasks, then present some experimental data, and finally we introduce a model which attempts to simulate feedback effects.

Only a few studies have examined the effect of feedback on production tasks. A larger number have used the 'start–stop' procedure developed by Kladopoulos et al. (1998; see Jazayeri & Shadlen, 2010; Rakitin & Malapani, 2008; Ryan & Fritz, 2007; Saito et al., 2015) which is sometimes described as a 'time production' task (e.g., by Rakitin & Malapani, 2008, and Saito et al., 2015), or even as 'estimation' (by Kladopoulos et al., 1998) but is in fact a type of reproduction. In a study of interval production per se, Montare (1985) asked people to produce 4 or 12 s by holding down a key under conditions without any feedback or with 'knowledge of results'. Data from males and females were analysed separately, and both groups showed increased accuracy with feedback. More striking, however, was a large reduction in variance in the feedback condition. A second study by Montare (1988) confirmed the effect that feedback improved mean accuracy on a 12-s production task (although the time produced was close to 12 s even without feedback). Once again, variance was reduced with feedback.

Wearden and McShane (1988) required people to produce intervals of 500, 700, 900, 1100 and 1300 ms, by pressing buttons to start and stop the interval. Feedback was provided after each trial. The relative frequency distributions of the times produced for each target time were constructed and fitted with Gaussian distributions (see Wearden and McShane's Figure 2, p. 369). The peak of these Gaussian curves (the mean) tracked target time very accurately with deviations from the target time of 20 ms at most (Wearden & McShane's Table 1, p. 370). The coefficient of variation of the curves (standard deviation/mean) was approximately constant. These data are shown in the present Fig. 5 as unconnected points.

Franssen and Vandierendonck (2002) reported data from both production and reproduction experiments, where the production task involved producing
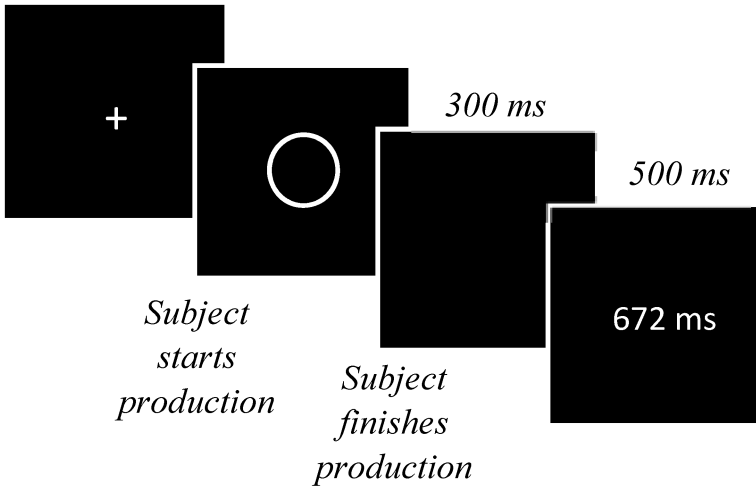
Figure 1. General outline of the temporal production task used here. In each trial, a fixation cross was shown until the participant initiated their production (spacebar). An empty circle was then presented until the participant terminated their production (spacebar). Feedback was presented for 500 ms displaying the duration the participant produced, followed by the next trial.
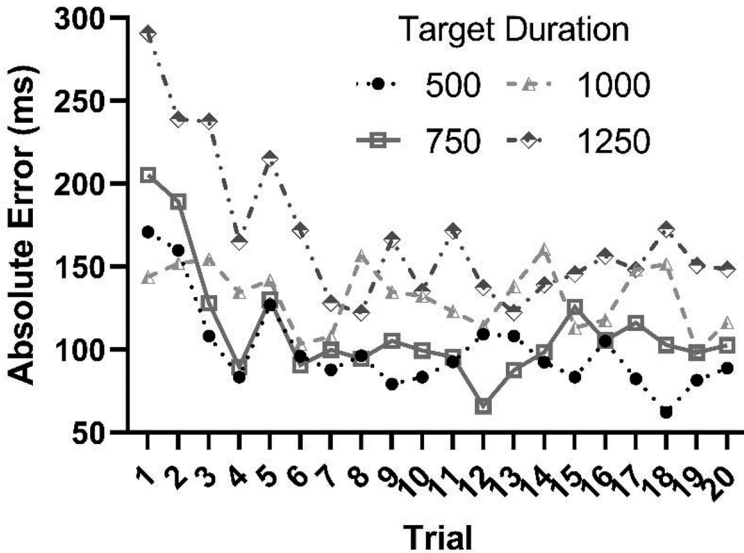
Figure 2. Mean absolute deviation from the target duration per trial, for all the target times.

intervals of either 4 or 12 s. Provision of feedback after a no-feedback phase resulted in greater timing accuracy with feedback than without.

Data from all these studies show that interval production with feedback can track target times very accurately on average, but the question of how feedback actually controls performance, and how this might be quantitatively modelled, has not previously been addressed.

We conducted an experiment where people were required to produce target intervals of 500, 750, 1000, and 1250 ms. These values were chosen because, like those used by Wearden and McShane (1988), they were probably too short to make chronometric counting useful to participants (Grondin et al., 1999). Because the experiment was run online, we wanted to test the reliability of the data obtained, so ran two identical experiments, identified as Experiments 1 and 2.

This article makes two main contributions to our understanding of temporal production. Firstly, we introduce several new trial-by-trial metrics to quantify performance. In particular, examining performance in terms of the number of times participants correctly adjusted their performance from one trial to another is revealing as to their use of the feedback provided. Secondly, this article provides a clear model which contributes to our understanding of how participants use feedback to perform temporal productions. Considering that the majority of experiments provide feedback either throughout all trials, or in a subset of them, understanding how feedback is incorporated into task performance is important.

2. Methods

2.1. Participants

All participants were gathered from a Psychology participant pool, and participation was in exchange for course credit. Participants provided digital consent, in accordance with the Declaration of Helsinki. These experiments were approved by the Macquarie University Ethics Committee.

Twenty-two participants were included in both experiments. In Experiment 1, four participants were replaced due to mean productions which were not monotonically increasing as a function of target time. Three participants were replaced in Experiment 2 for the same reason. The mean age of participants in Experiment 1 was 19.3 years [standard deviation (SD) = 2.4], 16 were female, and three were left-handed. The mean age of participants in Experiment 2 was 20.6 years (SD = 3.9), 18 were female, and two were left-handed.

2.2. Presenting software

Data was gathered online using the Gorilla experiment builder platform (gorilla. sc). See Anwyl-Irvine et al. (2020a) for an introduction to this online tool. Briefly,

this platform allows participants to perform pyschophysical experiments in their own home while gathering important metrics such as response times and choices, as well as system information such as operating system and browser details. Particularly in the visual modality, the timing of Gorilla has been found to be comparable to in-person data collection (Anwyl-Irvine et al., 2020b).

2.3. Procedure

To test data reliability because duration production experiments have not been performed using online methods, particularly Gorilla, the procedure was run twice in identical form. Note also that the size of the stimuli could vary, depending on the screen used by the participant.

At the beginning of the experiment, participants were given detailed instructions about what would occur on each trial. They then performed five practice trials (as described below) with a 500-ms target duration, which were not included in the analysis. Following this, participants performed four blocks of 20 trials each. In each block, they were required to produce either 500 ms, 750 ms, 1000 ms or 1250 ms. The target duration was told to the participant at the beginning of each block. The order of the blocks was randomized for each participant. The experiment took around 15 minutes to complete.

At the beginning of each trial, a fixation cross ('+') was presented centrally. The participant then initiated their production by pressing the spacebar once. Once the spacebar was pressed, the fixation cross was replaced with an empty circle, indicating that their production had begun. Participants then terminated their response by pressing the spacebar again. A 300 ms blank screen was then shown, followed by a screen displaying the duration the participant actually produced (in ms) for 500 ms. The next trial then began (see Fig. 1 for a schematic of the procedure).

2.4. Analysis

The analysis was done in several steps. Firstly, we calculated the mean, standard deviation and coefficient of variation (CV: SD/Mean) of participant productions for each participant in each experiment. Secondly, we calculated the amount each participant's productions changed from one trial to the next (current production − prior production), and calculated the mean, and SD of these changes, based on absolute values. Finally, we coded whether the current production was in the correct direction (i.e., towards the target duration) compared to the prior production. A correct direction was coded as a 1, and an incorrect direction as a 0. When averaged across the durations and participants, this gave a proportion of correct directional changes from the prior to current trial. On these mean measures, we performed an analysis of variance (ANOVA) with the target duration as a within-subjects factor, and the experiment as a between-subjects factor.

Finally, we fit a cognitive model to the combined data. This model was fit to four metrics; the mean duration produced, the CV of durations produced, the mean absolute production change from one trial to the next, and the SD of the changes from one trial to the next. The model was fit to the mean data from the group. Additionally, we examined whether the proportion of correct directions from prior to current trials was similar to that found in the data.

3. Results

The average trial by-trial productions are shown in Fig. 2. Graphics for the statistical analysis of productions are presented at the end of this section (Fig. 3). The top left displays mean productions for both experiments, and their average, the top right displays CVs, the bottom left shows mean per trial changes, the bottom middle shows per trial SDs, and the bottom right shows per trial directional change proportions.

3.1. Mean duration produced

The mean duration produced was significantly affected by the target duration ($F_{3,126} = 807.7$, Greenhouse–Geisser (GG)-corrected, $p < 0.001$, $\eta p^2 = 0.95$). The mean duration produced when data from the two experiments were combined for
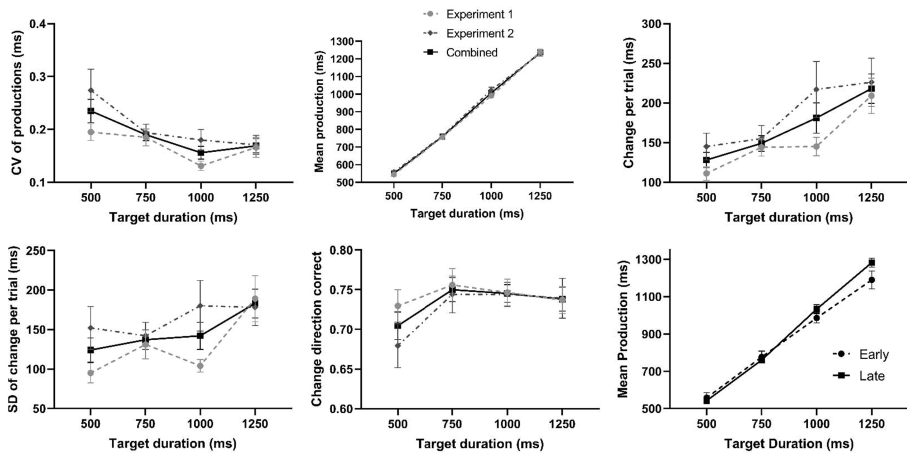
Figure 3. Data from Experiments 1 and 2, and their data combined. Upper left panel: Coefficient of variation (standard deviation/mean) of times produced. Upper middle panel: Mean time produced in ms. Upper right panel: Mean absolute change (ms) from one trial to another. Lower left panel: Standard deviation of absolute change (ms) from one trial to another. In all panels the measure illustrated is plotted against target time in ms. Lower middle panel: The proportion of trials in which participants correctly accounted for the prior offset by moving their production towards the target duration, in relation to their prior production. Lower right panel: Mean durations produced in the first three and last three trials at each target duration.

the 500-ms target was 551 ms, for the 750-ms target the mean produced duration was 759 ms, for the 1000-ms target the mean duration produced was 1007 ms, and for the 1250-ms target the mean production was 1236 ms. This effect of target time on average production was expected and not analysed further.

Neither the experiment ($F_{1,42} = 0.993$, $p = 0.325$, $\eta p^2 = 0.02$), nor the interaction of experiment and duration ($F_{3,126} = 0.640$, $p = 0.542$, GG-corrected, $\eta p^2 = 0.02$) were significant.

3.2. Mean CV

Mean CVs were significantly higher at shorter durations ($F_{3,126} = 7.32$, $p < 0.001$, GG-corrected, $\eta p^2 = 0.15$). The CVs for the 500-ms target (0.235) were significantly higher than at either the 1000 ms (0.156; $t_{43} = 3.99$, $p = 0.002$, Holm-corrected, $d = 0.58$) or 1250 ms (0.169; $t_{43} = 2.85$, $p = 0.034$, Holm-corrected, $d = 0.52$) target, but not compared to the 750-ms target (0.190; $t_{43} = 1.82$, $p = 0.227$, Holm-corrected, $d = 0.37$). No other condition comparison was significant (minimum p; $t_{43} = 2.30$, $p = 0.106$, Holm-corrected, $d = 0.43$).

The main effect of the experiment ($F_{1,42} = 3.13$, $p = 0.084$, $\eta p^2 = 0.07$), and the interaction between experiment and target duration ($F_{3,126} = 1.91$, $p = 0.149$, GG-corrected, $\eta p^2 = 0.04$) were both not significant.

3.3. Per trial change mean

The mean duration change from one trial to the next significantly varied by the target duration ($F_{3,126} = 9.87$, $p < 0.001$, GG-corrected, $\eta p^2 = 0.19$), with longer target durations resulting in larger mean changes. Mean changes were significantly higher given a 1250-ms target (218 ms) compared to either a 750-ms target (149 ms; $t_{43} = 3.82$, $p = 0.002$, Holm-corrected, $d = 0.66$) or a 500-ms target (128 ms; $t_{43} = 5.13$, $p < 0.001$, Holm-corrected, $d = 0.87$), but not in comparison to a 1000-ms target (181 ms; $t_{43} = 2.14$, $p = 0.114$, $d = 0.29$). Further, a 1000-ms target results in significantly larger mean changes compared to a 500-ms target ($t_{43} = 2.85$, $p = 0.027$, Holm-corrected, $d = 0.51$). No other comparison was significant (minimum p-value: $t_{43} = 1.73$, $p = 0.182$, Holm-corrected, $d = 0.33$).

The main effect of the experiment ($F_{1,42} = 2.62$, $p = 0.113$, $\eta p^2 = 0.06$), and the interaction between experiment and target duration ($F_{3,126} = 1.19$, $p = 0.313$, GG-corrected, $\eta p^2 = 0.03$) were both not significant.

3.4. Per trial change SD

The mean SD of trial-to-trial production changes was significantly affected by the target duration ($F_{3,126} = 3.49$, $p = 0.018$, $\eta p^2 = 0.08$). Generally, it appears that the SDs increased with increasing target durations (500 ms = 124 ms, 750 ms = 137 ms, 1000 ms = 142 ms, 1250 ms = 183 ms), however after Holm-correction, only the difference between the 500 ms and 1250-ms target was significant ($t_{43} = 2.84$, $p = 0.041$, Holm-corrected, $d = 0.53$). No other comparisons

were significant (minimum p-value $= t_{43} = 2.20$, $p = 0.166$, Holm-corrected, $d = 0.44$).

The main effect of the experiment ($F_{1,42} = 2.59$, $p = 0.115$, $\eta p^2 = 0.06$), and the interaction between experiment and target duration ($F_{3,126} = 2.19$, $p = 0.092$, $\eta p^2 = 0.05$) were both not significant.

3.5. Per trial directional changes

Neither the main effect of experiment ($F_{1,42} = 1.05$, $p = 0.311$, $\eta p^2 = 0.02$) or target duration ($F_{3,126} = 1.91$, $p = 0.131$, $\eta p^2 = 0.04$), nor the interaction ($F_{3,126} = 0.646$, $p = 0.587$, $\eta p^2 = 0.02$) reached significance. The mean proportion of correct directional changes was 0.734 (with 1 indicating always correct directional changes).

3.6. Improvement over trials

As can be seen in Fig. 2, participants appear to improve their productions dramatically from the first few trials to the final trials at all target durations except perhaps for the 1000-ms target duration. Comparing the mean duration produced over the second, third and fourth production (the first three productions after the first, in order for feedback to be available), and the mean durations produced over the 18th, 19th and 20th production showed that at the earlier target durations (500 ms and 750 ms), the mean durations produced did not significantly change ($t_{43} = 0.749$, $p = 0.916$, Holm-corrected, $d = 0.15$; $t_{43} = 0.647$, $p = 0.916$, Holm-corrected, $d = 0.139$). At the longer target durations (1000 ms and 1250 ms), the mean duration produced was closer in the earlier trials for the 1000-ms target (early $= 986$ ms, late $= 1033$ ms; $t_{43} = 2.76$, $p = 0.030$, Holm-corrected, $d = 0.40$), while the mean duration produced was closer to the target in the later trials for the 1250-ms target (early $= 1189$ ms, late $= 1282$ ms, $t_{43} = 2.81$, $p = 0.030$, Holm-corrected, $d = 0.50$).

Of note however, with the exception of both the early and late 500-ms trials (mean $= 557$ ms, $t_{43} = 2.93$, $p = 0.005$, Holm-corrected, $d = 0.44$; mean $= 542$ ms, $t_{43} = 5.09$, $p < 0.001$, Holm-corrected, $d = 0.77$, respectively), no other mean production was significantly different from the target (minimum p-value at 1000-ms target final trials; mean $= 1033$ ms; $t_{43} = 1.91$, $p = 0.124$, Holm-corrected, $d = 0.29$).

4. Modelling feedback effects on the production of short time intervals

Our modelling not only attempted to simulate overall performance in terms of mean productions and coefficients of variation but also attempted to understand how people are using the post-response feedback they were provided to guide their performance. In the experiments described earlier, exact numerical feedback was provided to participants, so not only did they know whether their previous production had been greater than or less than the target time, but they also

knew how much it had deviated from the target. It seems highly likely that this information was used to guide performance, but how? Inspection of a sequence of production trials gives the distinct impression of some kind of guided change from one interval to another, but variability in performance obscures any immediately clear picture. For example, a person may produce a value above the target, followed by another value also above the target, and not necessarily a closer one, in spite of receiving feedback that their first production was longer than the target time. However, it is possible that random variability from one trial to another contributes to this, rather than any failure to take note of the feedback delivered.

Feedback in interval timing has received little theoretical attention in the interval timing literature, although Franssen and Vandierendonck (2002) discuss how feedback might operate in the framework of the clock-counter model of scalar expectancy theory (Gibbon et al., 1984). They did not, however, attempt any quantitative modelling of the feedback effects obtained in the reproduction and production tasks they used.

The present model was developed using several assumptions. The first was that the production of a time interval involved a mixture of timing and non-timing (probably motor or motor preparation) processes, an idea that can be traced back at least to Wing and Kristofferson (1973). The second was that the non-timing process was invariant with respect to the interval timed, but the timing process obeyed scalar timing, that is, it had a mean which varied accurately on average with the interval to be timed and a constant coefficient of variation as the interval to be timed changed, as has been found to be the case for many different tasks where short time intervals are involved (Wearden, 1991; Wearden & Bray, 2001; Wearden & Lejeune, 2008). The third assumption was that responding from trial-to-trial would change depending on the feedback provided. We incorporated this idea by assuming that feedback could result in two processes. In one of these (adjust) the timing component of the trial subsequent to the feedback was adjusted by an amount on average equal to the discrepancy between the production on the previous trial and the target time. In the other (repeat) the production on the previous trial was repeated. The adjust process was invoked when the absolute discrepancy between the time produced on the previous trial and the target time exceeded some threshold. When the discrepancy was below the threshold the repeat process was invoked. The distinction between the two processes was designed to capture the intuitively reasonable supposition that people might not attempt to 'correct' productions when these are very close to the target time.

More formally, the basic structure of a trial was

$$P(n) = D + T$$

where P is the production on some trial n, D the duration of the non-timing process, and T the duration of the timing process on that trial. Depending on feedback, T changed from trial-to-trial, but D varied only randomly (see below). If X is the target time on that trial then the adjust process was

$T(n + 1) = T(n) + [X − P(n)]$

$T(n + 1)$ was then transformed into a value t^*, which was randomly drawn from a Gaussian distribution with a mean of $T(n + 1)$ and a CV of *cva*.

If the repeat process operated then $T(n + 1) = T(n)$, where $T(n + 1)$ was then transformed as above into a value t^*, a value drawn from a Gaussian distribution with mean $T(n + 1)$ and coefficient of variation *cvr*, and we assumed that *cvr* < *cva*, that is, the repeat process was less variable than the adjust process.

The repeat process operated if

$abs[X − P(n)] < b^*$,

and the adjust process if

$abs[X − P(n)] > b^*$

where b^* is a random value drawn from a Gaussian distribution with mean B (the mean threshold) and a CV of 0.5, and abs is absolute value.

$P(n + 1)$ was then constructed as

$P(n + 1) = t^* + d^*$,

where d^* is drawn from a Gaussian distribution with mean D, the mean of the non-timing process and a CV of 0.25.

As mentioned above, the model uses some ideas from previous models of interval timing and repetitive tapping. Along with the idea that an overt response resulted from a mixture of timing and non-timing processes (Wing & Kristofferson, 1973), we incorporated a simple feedback rule, namely that the adjustment on a trial was related to the discrepancy between the response produced and the target on the previous trial (similar to the adjustment used in Michon, 1967). The model also used the idea that timing decisions (in the present case the difference between adjust and repeat responses) are based on a comparison involving some kind of threshold, an idea common to several timing models, but particularly Wearden's (1992) model of temporal generalization performance. In the model the shortest production allowed was 250 ms, and the model would have resampled the trial had values less than this occurred, but they never did with the parameters used in the simulations reported.

The model was embodied in a Python program, and simulated values shown in Figs 4, 5, and 6 were derived from 10,000 trials. Our initial aim was to fit the four different data sets derived from the production experiment, the mean time produced, the CV of the time produced, the mean absolute change from one trial to another, and the standard deviation of the absolute change. This produced obvious problems of fitting, particularly as the different data sets were all on different scales. We endeavoured to keep as many parameters constant as possible between simulations. In particular, the CVs of the non-timing process and the threshold were always 0.25 and 0.5, respectively, the latter value being the same as in temporal generalization simulations (e.g., Droit-Volet et al., 2001).

Our fitting strategy was to try to find the minimum deviation between the output of the model and data for all four data sets at once, taking into account the
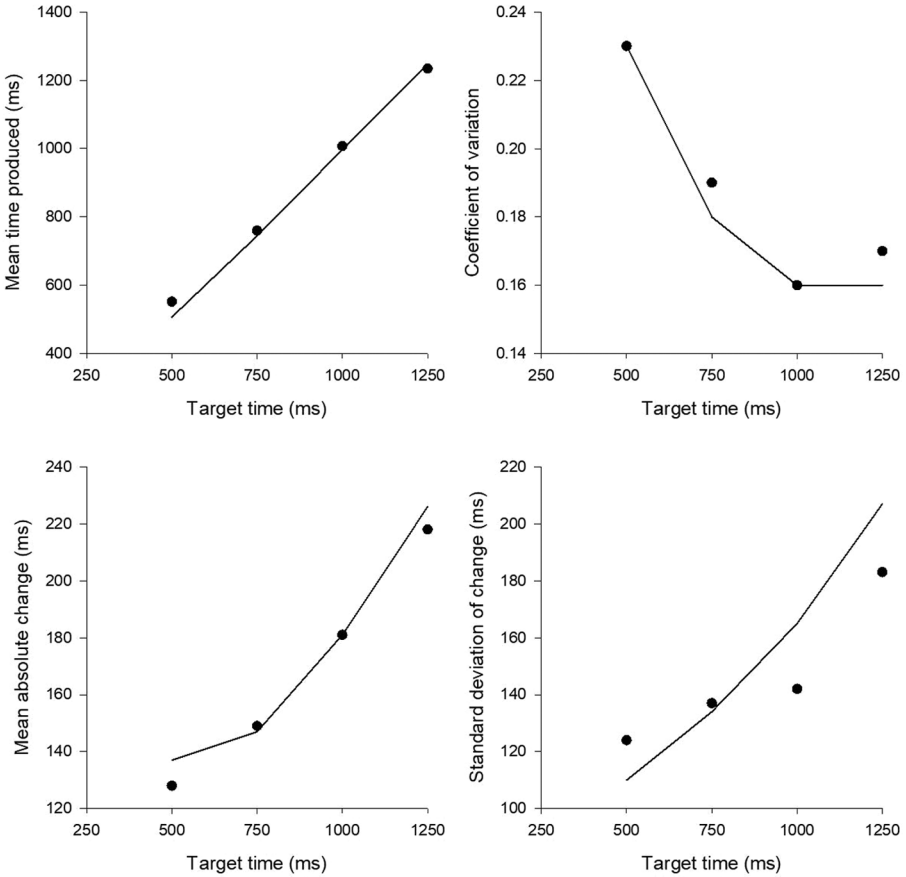
Figure 4. Top left panel: Mean times produced (ms). Top right panel: Coefficient of variation (standard deviation/mean) of times produced. Lower left panel: Mean absolute trial-by-trial change (ms). Lower right panel: Standard deviation of absolute trial-by-trial change (ms). In all panels the measure is plotted against target time, and is shown as unconnected points. The solid line within each panel shows results from the simulation discussed in the text.

different scales used for the different measures. When fitting the four different target times, the threshold was allowed to vary as a function of target time, but the coefficients of variation of the adjust and repeat processes, and the means of the non-timing process and the threshold, were kept constant.

Exploration of the model suggested that many parameter sets could fit the mean times produced well, and that the declining coefficient of variation as a function of target time could also be simulated by many parameter sets. The measures of change, however, were more problematic, particularly the standard deviation of the change. Eventually, we arrived at a compromise parameter set which resulted in the simulation values shown in Figure 4 as lines in each panel. The
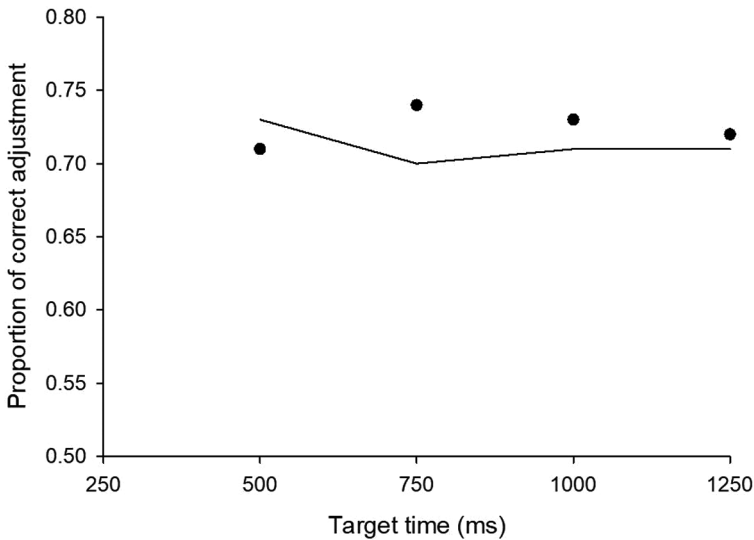
Figure 5. Directional change measure discussed in the text. Unconnected points show data from the combined data set. The solid line shows the model predictions.

compromise resulted in fits for each individual measure that were poorer than could be obtained if that measure were the only focus of the simulation.

For the simulations shown in Figure 4, $D = 122$ ms, $cvr = 0.11$, and $cva = 0.14$. The mean threshold B varied with target time, and took values of 120, 160, 160, and 170 ms for the 500-, 750-, 1000-, and 1250-ms target times, respectively.

In Fig. 4, the top left panel shows fits to the mean times produced, and the top right one fits to the coefficients of variation. The bottom panels show fit to the mean of the absolute trial-by-trial change (bottom left) and the standard deviation of the change (bottom right). In qualitative terms, the simulation fitted the data well in all cases. Predicted means increased, and coefficients of variation decreased, as the target time lengthened. For the two absolute change measures, predicted values increased with target time, as in the data.

Quantitatively, the simulation output fitted the mean times produced, the coefficients of variation, and the mean absolute change fairly well. The average percentage absolute deviation between the output of the model and data was 3%, 4% and 3%, respectively for the measures of mean time produced, coefficient of variation, and mean absolute trial-by-trial change. The standard deviation of the absolute change was fitted less well, with a mean percentage deviation of 11%. Exploration of the parameter values used in the model showed, in fact, that this measure was the most problematic one to fit in all the simulations we used.

Given that it seemed that the model did not capture all aspects of trial-to-trial change completely accurately, we simulated the directional changes in the data analysed above. This was the proportion of changes in the 'correct' direction: that
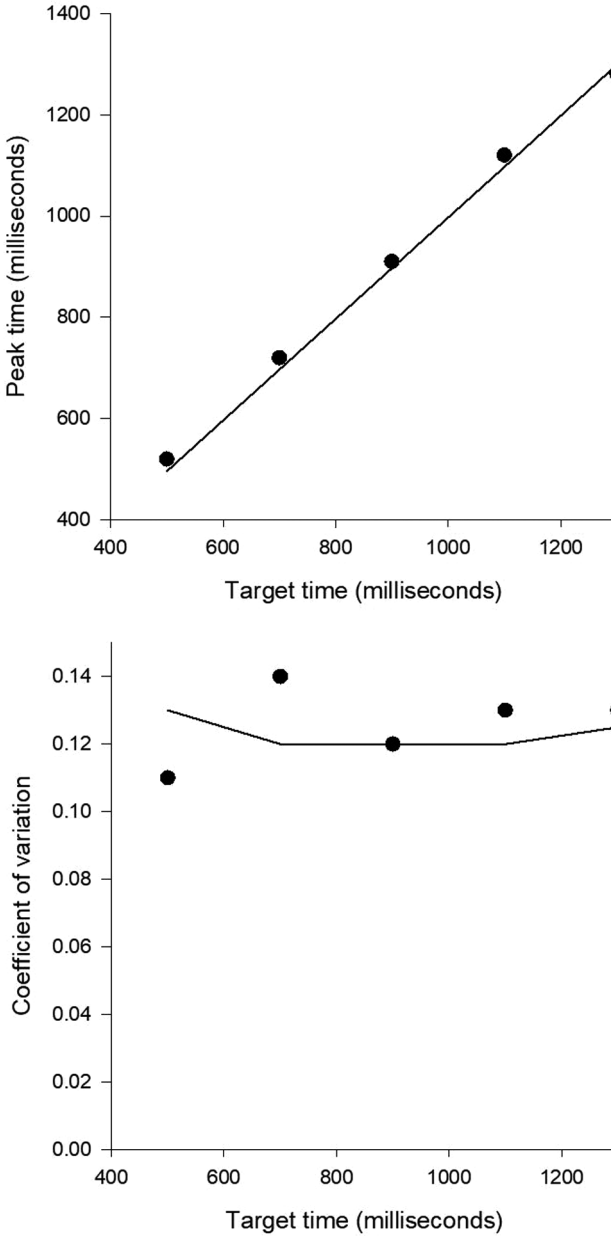
Figure 6. Data from Wearden and McShane (1988). Upper panel: peak time (ms) plotted against target time. Lower panel: Coefficient of variation plotted against target time. Unconnected point show data from Wearden and McShane (1988) Table 2. The line in each panel shows results from a simulation discussed in the text.

is, a decrease in the duration of the production on trial $N + 1$ if the production on trial N had been above the target time, and an increase in the duration of the production on trial $N + 1$ if the production on trial N had been below the target time. This measure gives a kind of non-parametric measure of trial-by-trial change. The model parameters were not adjusted specifically to simulate this measure, so the simulation has the status of a prediction. Figure 4 shows the data from the combined data set and the output of the model. It is clear from Fig. 4 that the model produced output very close to the data, even though no attempt was made to find the best-fitting parameters for this measure.

The constancy of some of the parameters in our model accorded with intuition. The relative variability of timing, including timing the change and repeat processes, should not vary as the interval timed varied, so a constant coefficient of variation was used, and this was found to fit data well. Likewise, the non-timing process, presumably reflecting motor or motor preparation processes, should not vary as a function of the interval timed when the response was constant, and that also was kept constant in the simulation. On the other hand, the threshold we used might be expected to vary with target time, as differences between time intervals are easier to discriminate at shorter intervals than longer ones, but in fact in our simulations the threshold increased by only 33% between the 500-ms and 1250-ms target times. A more elegant solution would have been to have a threshold which was a constant fraction of the target time, as this varied. This idea was tried in earlier versions of the model, but failed to produce good fits to the absolute change measures. To fit our data, the threshold had to be a slightly decreasing percentage of the target time, as this increased (e.g., from 24% at 500 ms to 13.6% at 1250 ms).

Some of the effects obtained by varying our simulation parameters are obvious: for example, increasing the CV of the adjust and repeat processes increased the variability of productions. Another effect which is obvious after a moment's reflection is that increasing the magnitude of the non-timing process increased the CVs of the times produced but, more importantly, made the CV as a function of target time decline more steeply. This occurs because the non-timing process, which is on average of constant duration, makes a progressively smaller contribution to the variance of productions as the target time increases.

To illustrate this, we attempted to simulate data from Wearden and McShane's (1988) study of the production of five time intervals from 500 to 1300 ms in length. At first sight, Wearden and McShane's data appear problematic for our model in that the coefficient of variation of the times produced was approximately constant, not declining with increasing target time as in the present study. In Wearden and McShane's experiment, people produced the time intervals by pressing different response buttons to start and stop the production and received immediate numerical feedback from the experimenter. The four participants received 120 trials at each target time rather than the 20 trials given in the present experiment. The numerical data from Wearden and McShane's study are no longer available,

but means and CVs of the times produced can be approximated using data in their Table 1 (p. 370). The measures in this Table come from fits of Gaussian curves to the relative frequency of times produced plotted against target time. This yielded a peak time (a measure of the mean time produced) and curve coefficient of variation. The values obtained as shown in Fig. 5 as unconnected points.

To simulate Wearden and McShane's data, we reduced the mean of the non-timing process to 50 ms, and set the threshold at 100 ms. The coefficients of variation of the repeat and adjust processes were 0.08 and 0.11, with other parameters having the same values as for the simulations shown in Fig. 4. Inspection of Fig. 6 suggests that the fit was generally good, with data and the simulation results showing approximately scalar timing (i.e., mean close to target time, and constant coefficient of variation). The ability of the model to fit data which shows a more or less constant coefficient of variation as the time produced varied, and the present data which show declining coefficients of variation show that the two types of data are not really incompatible. Although there were only four participants in Wearden and McShane's study, they did receive more extensive training than the participants in the current work, and the data seem reliable even at the individual level, as no participant showed a declining coefficient of variation with increasing target time (see Wearden and McShane, 1988, Table 1, p. 370).

A manipulation which has featured in several articles using the start–stop procedure has been the provision of 'false feedback'. That is, people were told that their response time was some percentage of its real value, such as 80 or 120% (Ryan & Fritz, 2007). The effect is usually to shift the response measure in the appropriate direction, so reproductions are longer in the 80% feedback condition, and shorter in the 120% one, than in the 100% condition, where the feedback is veridical. In our model, such false feedback would be tantamount to changing the target time and Fig. 5 shows the effect on this manipulation of the mean times produced when the feedback was 80, 100, or 120% of its veridical value, and the model produced target times of 500, 750, and 1000 ms. Our model would be expected to be sensitive to the false feedback manipulation, and it is, as Fig. 7 shows, with the mean times produced increasing or decreasing, depending on the false feedback, relative to the 100% condition where feedback is veridical. For the simulations in Fig. 6 all parameter values were as the Wearden and McShane (1988) simulation shown in Fig. 5, and the false feedback conditions were simulated by varying the effective target time.

5. Discussion

Our model breaks new ground in that it simulates, for what seems to be the first time, the effects of accurate numerical feedback on interval production, particularly from one trial to the next. Feedback or calibration in timing studies is often a given, not further discussed. For example, in stimulus timing tasks like temporal
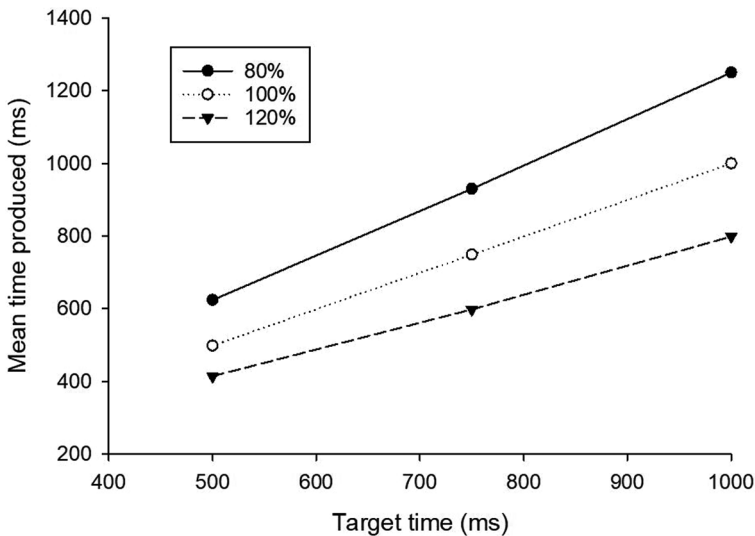
Figure 7. Simulation of 'false feedback' procedure where the feedback was 80, 100, or 120% of its real value. See text for other details.

generalization (Wearden, 1992) and temporal bisection (Wearden, 1991) standards must be initially presented so that comparison stimuli can be judged against them. A small number of studies have reported feedback effects on various tasks, for example verbal estimation (Montare, 1988; Wearden & Farrar, 2007), production (Franssen & Vandierendonck, 2002; Montare, 1985, 1988), reproduction (Franssen & Vandierendonck, 2002; Riemer et al., 2019; Ryan & Robey, 2002), and the start–stop procedure (e.g., Ryan & Fritz, 2007). Although various effects have been obtained, usually a reduction in variance compared with no-feedback conditions, and improved accuracy, the effects of feedback have not been quantitatively modelled for these tasks.

Our model grasped the nettle of attempting an explicit quantitative model of feedback but, it must be admitted, it was not a complete success. One issue is goodness of fit: our model does not fit the overall data set it simulates as well as models of, for example, temporal generalization (e.g., Wearden, 1992). However, in the cases where fit is much better, only a single outcome variable, usually response probability, is modelled, whereas here we attempt, more ambitiously, to model four measures of interval production which characterize both the mean output and aspects of its variability and trial-by-trial change. As mentioned earlier, if one of the outcome measures alone was the focus of the modelling, better fits could be obtained, but only at the cost of making the fit to the other measures worse. Throughout our attempts to model the data, the standard deviation of the absolute trial-by-trial change was always the most problematic to fit, although almost all parameter settings of the model produced a standard deviation which increased with the target time, often more sharply than found in data. Nevertheless, in spite

of the difficulty of simulating this particular measure completely accurately, the non-parametric measure of change was fitted fairly well without specific changes in the parameter set used, so some aspects of trial-by-trial change, such as mean absolute change, and the directional change, can be fitted well.

A second issue with our model is that it uses only the single previous production as the basis for performance on the current trial, and does not incorporate any longer-term representation of the target time. This means that it is unclear what it would predict were feedback removed. Adding the development of some sort of temporal reference to the model is a potential step, but we did not do this as it would complicate a model which already incorporated several different processes and which seemed to fit data reasonably well as it was. However, to simulate effects of removing feedback, or to simulate different types of feedback conditions, such as just 'above or below target' with no numerical information, or no feedback at all, other processes would need to be added.

Most of the parameters in our model have obvious psychological meanings: the adjust and repeat timing processes have associated variance, and a threshold determines which occurs on the next trial. These are similar in principle to the parameters used in models of stimulus timing, like temporal generalization (Wearden, 1992). However, the one parameter which is potentially more opaque is the non-timing process. We originally envisaged this as a response time, so our first idea was that a timing process would be followed by the generation of a response which would take some time to execute, and this execution time (supposed to be on average constant as target time varied) would be represented in the non-timing process. This may still be true, but in our simulation the mean of the non-timing process was shorter than any conceivable reaction time, so if it is related to motor output then it seems that the motor output preparation must be running at least in part in parallel with the timing process rather than coming after it. In our simulation realistic values for a response time, such as 250 ms or more, could not produce simulations in accord with the data.

As an alternative to motor delay, it is possible that this parameter represents motor variability. On average, people tend to be able to accurately time when to release a response so that its execution coincides with a given moment. This can be seen, for example, in the tapping literature; once the metronome stimulus has disappeared participants are still able to maintain a paced tapping speed relatively well, see Repp (2005) for review. In the duration production task here, it could be that there is some afferent (i.e., when timing starts versus when the button is actually pushed to start the production) and efferent (i.e., when timing finishes versus when the button push terminates the production) motor-timing variability. This duration is likely shorter than that of a reaction time-like interpretation. To test this, perhaps various methods of production could be used (e.g., termination only); however, these have associated difficulties, such as predicting when an automatically started production would begin (see Wehrman, 2020). Whatever the interpretation of this parameter, however, it is clear that a non-time-constrained

component must be present in order to account for the decreasing relative vari-ability as the target duration gets longer, which is generally not found in non-motor timing tasks (e.g., in temporal generalization, see Wearden, 1992).

Overall, our model incorporated a number of previous ideas in a simple way. Production output resulted from a combination of a timing and non-timing pro-cess (cf. Wing & Kristofferson, 1973). When adjustment between trials occurred, it was based on the discrepancy between the previous production and the target time (cf. Michon, 1967). Both timing and non-timing processes had associated variance, and a threshold is used to control performance (cf. Wearden, 1992). It may be that there are sources of variability in data that the model does not cap-ture completely accurately but, as Fig. 3 shows, it can produce output which is a reasonable approximation to all four of the data measures used, which themselves capture different aspects of performance on interval production.

6. Conclusion

In the current article, participants performed a simple duration production task. We found that, over relatively short time intervals, participants were able to cor-rectly produce target durations, and had decreasing variability as the target duration increased. In addition, we showed that trial-to-trial changes increased as the target duration increased, as did the variability of these changes. Further, participants were approximately 75% accurate in their use of feedback, such that their subsequent production moved towards the target duration in comparison to the prior produc-tion. We fit a model to these measures to capture how feedback affects participants on both an average and trial-by-trial basis. This model had three key components, a scalar timing process, a threshold for 'acceptable' productions leading to repeat or adjust processes, and a stable non-timing process. This model fit the data well, with perhaps the exception of trial-by-trial variability. However, this model represents a first step in a more thorough understanding of duration production, and the role of feedback in this process. Further studies could incorporate, for example, different types of feedback, or different methods of production.

References

Anwyl-Irvine, A. L., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2020b). Online Timing Accuracy and Precision: A comparison of platforms, browsers, and participant's devices doi: 10.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020a). Gorilla in our midst: An online behavioral experiment builder. *Behav. Res. Methods*, *52*, 388–407. doi: 10.3758/s13428-019-01237-x.

Droit-Volet, S., Clément, A., & Wearden, J. (2001). Temporal generalization in 3- to 8-year old chil-dren. *J. Exp. Child Psychol.*, *80*, 271–288. doi: 10.1006/jecp.2001.2629.

Franssen, V., & Vandierendonck, A. (2002). Time estimation: Does the reference memory mediate the effect of knowledge of results? *Acta Psychol.*, *109*, 239–267 doi: 10.1016/S0001-6918(01)00059-2.

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. *Ann. N. Y. Acad. Sci.*, *423*, 53–77 doi: 10.1111/j.1749–6632.1984.tb23417.x.

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 performance of humans. *J. Exp. Psychol. Hum. Percept. Perform.*, *25*, 993–1004 doi: 10.1037/0096-1523.25.4.993.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nat. Neurosci.*, *13*, 1020–1026 doi: 10.1038/nn.2590.

Kladopoulos, C., Brown, B. L., Hemmes, N. S., & Cabeza de Vaca, S. (1998). The start-stop procedure: Estimation. *Percept. Psychophys.*, *60*, 438–450 doi: 10.3758/BF03206865.

Michon, J. A. (1967). *Timing in temporal tracking*. Assen, The Netherlands: Van Gorcum.

Montare, A. (1985). Learning effects of knowledge of results upon time estimation. *Percept. Mot. Skills*, *60*, 871–877 doi: 10.2466/pms.1985.60.3.871.

Montare, A. (1988). Further learning effects of knowledge of results upon time estimation. *Percept. Mot. Skills*, *66*, 579–588 doi: 10.2466/pms.1988.66.2.579.

Rakitin, B. C., & Malapani, C. (2008). Effect of feedback on time production errors in aging participants. *Brain Res. Bull.*, *75*, 23–33 doi: 10.1016/j.brainresbull.2007.07.014.

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychon. Bull. Rev.*, *12*, 969–992. doi: 10.3758/BF03206433.

Riemer, M., Kubik, V., & Wolbers, T. (2019). The effect of feedback on temporal error monitoring and timing behavior. *Behav. Brain Res.*, *369*, 11929 doi: 10.1016/j.bbr.2019.111929.

Ryan, L. J., & Fritz, M. S. (2007). Erroneous knowledge of results affects decision and memory processes on timing tasks. *J. Exp. Psychol. Hum. Percept. Perform.*, *33*, 1468–1482 doi: 10.1037/0096-1523.33.6.1468.

Ryan, L. J., & Robey, T. B. (2002). Learning and performance effects of accurate and erroneous knowledge of results on time perception. *Acta Psychol.*, *111*, 83–100 doi: 10.1016/S0001-6918(02)00044-6.

Saito, C., Janssen, S. M. J., & Tayama, T. (2015). The effects of performance feedback on interval timing: Learning and generalization. *Jpn. J. Psychon. Sci.*, *34*, 27–34 doi: 10.14947/psychono.34.4.

Wearden, J., & Bray, S. (2001). Scalar timing without reference memory? Episodic temporal generalization and bisection in humans. *Q. J. Exp. Psychol.*: *54B*, 289–309 doi: 10.1080/713932763.

Wearden, J.H. (1991). Human performance on an analogue of an interval bisection task. *Q. J. Exp. Psychol.*, *43B*, 59–81 doi: 10.1080/14640749108401259.

Wearden, J.H. (1992). Temporal generalization in humans. *J. Exp. Psychol. Anim. Behav. Process.*, *18*, 134–144.

Wearden, J.H., & Farrar, R. (2007). Effects of feedback and calibration on the verbal estimation of the duration of tones. *Acta Psychol.*, *126*, 1–17 doi: 10.1016/j.actpsy.2006.10.003.

Wearden, J. H., & Lejeune, H. (2008). Scalar properties in human timing: Conformity and violations. *Q. J. Exp. Psychol.*, *61*, 569–587 doi: 10.1080/17470210701282576.

Wearden, J.H., & McShane, B. (1988). Interval production as an analogue of the peak procedure: Evidence for similarity of human and animal timing processes. *Q. J. Exp. Psychol.*, *40B*, 363–375. doi: 10.1080/14640748808402330.

Wehrman, J. (2020). Temporal productions in a variable environment: timing starts from stimulus identification rather than onset. *Psychol. Res.*, doi: 10.1007/s00426-020-01430-0.

Wing, A. M., & Kristofferson, A. B. (1973). The timing of interresponse intervals. *Percept. Psychophys.*, *13*, 455–460 doi: 10.3758/BF03205802.