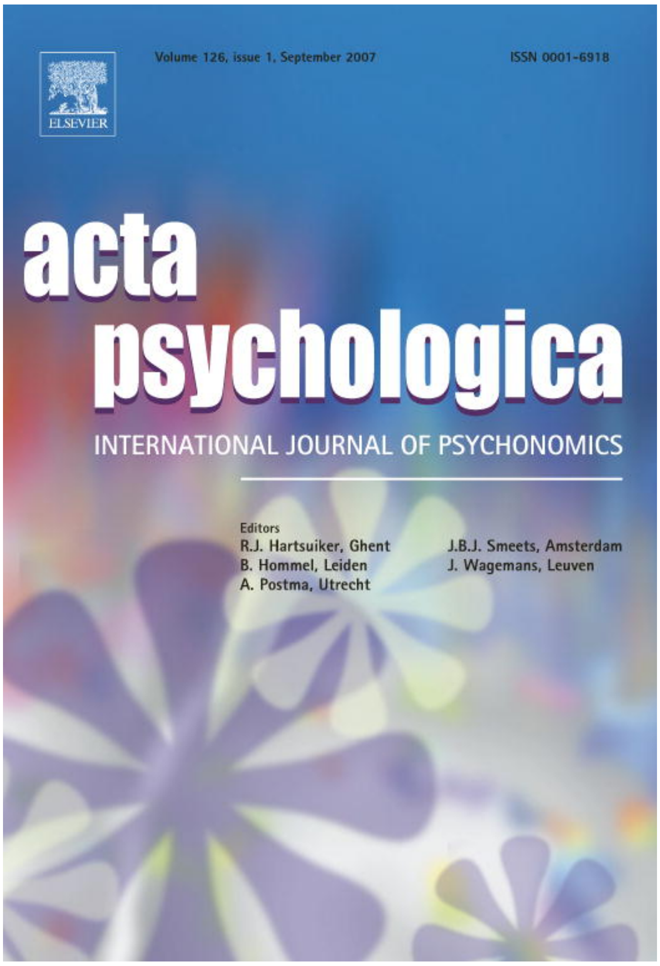
Volume 126, issue 1, September 2007
ISSN 0001-6918
ELSEVIER
acta psychologica
INTERNATIONAL JOURNAL OF PSYCHONOMICS
Editors
R.J. Hartsuiker, Ghent
B. Hommel, Leiden
A. Postma, Utrecht
J.B.J. Smeets, Amsterdam
J. Wagemans, Leuven

Available online at www.sciencedirect.com
ScienceDirect
Acta Psychologica 126 (2007) 1–17
acta psychologica
www.elsevier.com/locate/actpsy

# Effects of feedback and calibration on the verbal estimation of the duration of tones ☆

J.H. Wearden *, Rebecca Farrar

*School of Psychology, Keele University, Dorothy Hodgkin Building, Keele, Staffordshire ST5 5 BG, UK*

Received 15 November 2005; received in revised form 2 October 2006; accepted 28 October 2006
Available online 6 December 2006

**Abstract**

Two experiments investigated the effects of feedback (Experiment 1) and calibration (Experiment 2) on verbal estimation of the duration of tones ranging from 77 to 1183 ms in length. In Experiment 1, accurate feedback was provided after half the trials in the feedback condition, and this was contrasted with a no feedback condition in which feedback was absent. In Experiment 2, in the calibration condition, participants estimated the duration of the second of two tones, after a first tone of a known duration (ranging from 50 to 1200 ms) had been presented. In the no calibration condition the value of the first tone was unknown. Compared with conditions without feedback or calibration, the feedback/calibration operation produced no significant change in group mean estimates, nor coefficients of variation of estimates, but the absolute deviation of estimates from the real stimulus duration, and the absolute deviation of individual estimates from the group mean, were both significantly reduced. The feedback/calibration thus made participants' mean estimates thus more accurate and more similar to one another, but the actual mechanism by which this was done was apparently a subtle process of adjustment, rather than a gross change in mean estimate or estimate variability.

*PsycINFO classification:* 2340

*Keywords:* Time perception; Verbal estimation; Feedback; Calibration; Scalar timing

☆ The second author conducted some of the, experimental work reported in this article, while an undergraduate student at Manchester University.

* Corresponding author.
*E-mail address:* j.h.wearden@psy.keele.ac.uk (J.H. Wearden).

0001-6918/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.actpsy.2006.10.003

## 1. Introduction

The verbal estimation of duration, where people assign verbal labels, in conventional time units such as seconds or milliseconds, to the duration of events they experience, is at once one of the most useful, and one of the most mysterious, of all the techniques available for the study of the perception of time. Although verbal estimation has been used since the earliest days of time psychology (see Fraisse, 1964, for a historical account), recent studies have employed it fruitfully to investigate properties of the putative "internal clock" in humans. For example, Wearden, Edwards, Fakhri, and Percival (1998) proposed that the well-known auditory/visual difference in duration perception, with auditory stimuli being judged as longer than visual ones, is due to different "clock speeds" in the two modalities: that is, the pacemaker of a hypothesised pacemaker-accumulator clock (Gibbon, Church, & Meck, 1984) was proposed to run at a faster rate for auditory than for visual stimuli. A prediction of this "clock speed" hypothesis is that the auditory/visual duration perception difference should increase with increases in absolute duration, and to test this as large a range of duration values as possible should be used. Verbal estimation provided an apparently ideal method to test this idea. If the focus of interest is timing without the intervention of chronometric counting, then the use of durations less than about 1200 ms is advisable (Grondin, Meilleur-Wells, & Lachance, 1999), and methods like production or reproduction of time intervals are necessarily restricted by reaction time considerations as to the shortest duration that can be studied. Using verbal estimation, on the other hand, very short durations can be used, so the ratio of the longest (around 1200 ms) to the shortest can be 15 or 20 without methodological difficulty. Wearden et al. (1998) found support for the "clock speed" interpretation of auditory/visual differences using verbal estimation (for further work on this issue see Wearden, Todd, & Jones, 2006). For other studies of "clock speed" exploiting verbal estimation see Penton-Voak, Edwards, Percival, and Wearden (1996).

On the other hand, although verbal estimation is a very useful technique, it is also mysterious in the sense that we have little or no idea about the detailed mechanisms by which people perform on the task, such as potential "scales" they may be using to generate their responses, how "raw" stimulus duration representations are translated into the numbers the participants produce, and so on. In the studies cited above (Penton-Voak et al., 1996; Wearden et al., 1998) this ignorance was not a real problem, as verbal estimation of the duration of stimuli of one type was contrasted with verbal estimation of duration of stimuli of another type, and there was no reason whatsoever to suppose that the processes underlying verbal estimation (whatever they are) differed from one stimulus type to another.

Nevertheless, although verbal estimation can be used in many cases without a detailed knowledge of its psychological mechanisms, uncovering these psychological mechanisms would not only aid interpretation of experimental data, but would also contribute to the theoretical coherence of the field of time perception. An important recent development in the study of timing has been the application of Gibbon et al.'s (1984) scalar timing (or scalar expectancy) theory, SET, to timing in humans (see Allan, 1998; Wearden, 2003, for reviews). SET exists in a number of forms, but one of these allows the development of quantitative models of performance on timing tasks, and these have been found to fit data from experiments on human timing to a high degree of precision (see Wearden, 1992; see also Allan & Gibbon, 1991; for a survey of timing models see Wearden, 2004). However, as

Wearden (2003) points out, SET has received relatively little application to the "classical trio" of timing tasks, production, reproduction, and verbal estimation of duration, and has instead been mostly applied to tasks like bisection (Wearden, 1991) and temporal generalization (Wearden, 1992), which relate to procedures used with animals, to which SET was originally applied.

Wearden (2003) and Wearden and Lejeune (submitted for publication) discuss the application of SET to "classical" timing tasks, and conclude that, in general, deviations from the predictions of SET are quite common when these tasks are used. For example, a requirement of SET is "mean accuracy", the requirement that average "estimates" of some real duration, *t*, should vary linearly and accurately with changes in *t*. Classical timing tasks violate this assumption, usually by deviating in the direction of *Vierordt's Law*, the assertion that short durations are overestimated, whereas longer ones tend to be more accurately estimated or underestimated. The second requirement of SET is the "scalar property" of variance, the requirement that the standard deviation of time judgements should be a constant fraction of the mean, or equivalently that the coefficient of variation of time judgements (standard deviation/mean) should remain constant as the interval timed varies. Wearden (2003) showed data suggesting that production, reproduction and verbal estimation violate this scalar property, something which, to anticipate results to be presented later, is also true of the verbal estimation data collected in the present work.

One possibility is that such violations of SET reflect fundamental differences in the way that time is perceived in tasks like temporal generalization (where behaviour conforms to the requirements of SET) and verbal estimation (where behaviour violates SET), but this seems extremely unlikely given that the stimuli to be judged are often the same or very similar in the two types of tasks. It seems much more likely that the apparent violations of SET in classical tasks are due to as yet unknown factors involved in the performance of such tasks, but research to date has given us little indication of what these factors might be.

The present article contributes to the understanding of verbal estimation by presenting data from experiments in which participants received either feedback, or calibration, in some conditions where the duration of tones were estimated. A common technique (Wearden et al., 1998) is to present participants with short-duration tones, and ask for estimates in ms (i.e. using a scale where 1000 = 1 s). Usually, neither pretraining nor feedback is given to participants, who often complain that they have "nothing to go on" when making their estimates. This does not prevent their estimates from being orderly (e.g. Penton-Voak et al., 1996), but raises the obvious question of just what feedback or "calibration", which here means giving participants examples of stimuli of specified duration, would do to performance.

The study of how feedback or various sorts of calibration operations influence performance on timing tasks potentially enables us to understand something about how these tasks are performed, in particular, to understand something about the "standards" that people might be using to make their judgements. In addition, understanding how people use feedback may give us some insight into underlying psychological mechanisms involved in time judgements, such as the memory and decision processes used. For example, Droit-Volet and Izaute (2005) used a temporal generalization task with and without feedback, and participant groups of 5- and 8-year-old children, and adults. In temporal generalization in humans, people are initially given a standard stimulus duration to remember (in Droit-Volet & Izaute's study this was a tone 600-ms long), then receive durations longer than, shorter than, or equal to the standard and have to judge whether or not each one was or was not the standard (making a YES/NO response). A temporal generalization gradient can be drawn using the proportion of YES responses made at each stimulus value plotted against stimulus duration: the location of the peak of this gradient and its spread indicate, respectively, accuracy and variability of temporal judgements.

Droit-Volet and Izaute (2005) found that feedback made judgements less variable, although gradients peaked at the standard duration even without feedback, and computer simulation of the data using a model developed by Wearden (1992) suggested that the feedback made the representation in memory of the standard (600-ms) duration more precise, as well as reducing the tendency of the younger children to respond at random without regard to stimulus duration. However, there was a clear tendency for the no feedback/feedback difference to diminish with increasing age, with adults showing the smallest difference between conditions. Franssen and Vandierendonck (2002) studied effects of feedback on either reproduction or production of multi-second time intervals in adults. Feedback made reproductions longer, closer to the target time, and more variable when participants were instructed to divide their attention between visual and temporal features of the stimuli reproduced, but in general effects were small. On the other hand, production of time intervals of 4 and 12 s was substantially affected by feedback, with times produced being closer to target times and less variable with feedback than without. Franssen and Vandierendonck (2002) conclude that the main effect of feedback is on the "reference" that people are using on the task. One of the issues of interest in verbal estimation is the general question of the "reference" that people use to perform and, as in the studies discussed above, potential effects of feedback may offer some insight into this.

## 2. Experiment 1

Experiment 1 introduces a technique for providing feedback in a verbal estimation task. A moment's thought suggests that the question of how feedback might be provided in verbal estimation has no immediately obvious answer. In an interval production task, the participant attempts to produce some target time, *t*, then can receive feedback as to the time actually produced, then another trial with the same target time, and so on (see Wearden & McShane, 1988; Wearden, Wearden, & Rabbitt, 1997, for data obtained with this method). However, we cannot present a participant with some stimulus, tell the person what its duration actually was, then present the stimulus again on the next trial. Even intermixing a number of different durations might cause difficulties, as the participants may quickly learn that only a small number of duration values are being presented, and learn to identify them.

Experiment 1 avoids these problems as follows. The participant received a series of tones, and was required to estimate the duration of each one, using a scale where 1000 = 1 s. A random half of the tones were followed by accurate feedback, where the participant was told what the duration (in ms) of the stimulus just presented was, after making his/her own estimate. The other half of the tones were not followed by feedback, and the two tone types were randomly intermixed. In fact, the two types of tones were different. The tones after which feedback was given had previously been randomly generated from a range running from 50 to 1500 ms, and their actual range was from 97 to 1496 ms. These tones served to potentially "calibrate" the participant but were not otherwise of interest. The tones after which no feedback was given were repeats of one of 6 tone durations, ranging from 77 to

1183 ms, and it is the estimates of the duration of these tones which were the focus of the study. A condition in which feedback was provided for the "calibration" tones was contrasted with one where the same stimuli are presented, but without any feedback.

## 2.1. Method

### 2.1.1. Participants

25 Manchester University Psychology Department undergraduates participated for course credit.

### 2.1.2. Apparatus

An IBM-compatible PC controlled all experimental events. The stimuli to be judged were 500-Hz tones produced by the computer speaker, and the computer keyboard registered responses. The program running the experiment was written in MEL (Micro-Experimental Laboratory: Psychology Software Tools).

### 2.1.3. Procedure

All participants received two experimental sessions, one with feedback and the other without. Twelve participants received the feedback session first, the other 13 the no-feedback session first.

The feedback session consisted of 72 trials. For 36 of these, the stimulus durations were 77, 203, 461, 767, 958, and 1183 ms, with each value being presented 6 times. No feedback was ever given after these stimuli had been presented and their durations estimated. The other 36 trials of the session consisted of the presentation of 36 other "calibration tones" with durations ranging from 97 to 1496 ms. The durations had been previously generated by randomly sampling from a uniform distribution running from 50 to 1500 ms, but the same values were used for each participant. To produce each stimulus to be judged, the participant responded after a "Press spacebar for next trial" prompt, and this was followed after a random delay ranging from 2000 to 3000 ms by the stimulus presentation. Offset of the stimulus was followed by a prompt asking participants to type in their estimate. Participants had previously been asked to use a scale where "1000 = 1 s". They were also informed that all stimulus durations were between 50 and 1500 ms, and were asked not to use estimates outside this range. When the estimate had been typed in, the participant either received the "Press spacebar" prompt again (after no-feedback trials), or received the following display for 2 s: "The duration of the tone you just heard was*X*milliseconds", where *X* was the actual duration of the tone, after the "calibration tone" trials. The information given to participants was always correct. When the display terminated, the participant received the "Press spacebar" prompt. The 72 trials were arranged in a random order which was different for each participant. The no-feedback session was identical in all respects, apart for appropriate changes in experimental instructions, except that all 72 trials were no-feedback trials.

## 2.2. Results

Fig. 1 shows data from the no feedback (NF) and feedback (F) conditions of Experiment 1. The upper panel shows mean estimates, and the lower panel coefficients of variation of estimates (standard deviation/mean), for the 6 stimulus durations presented

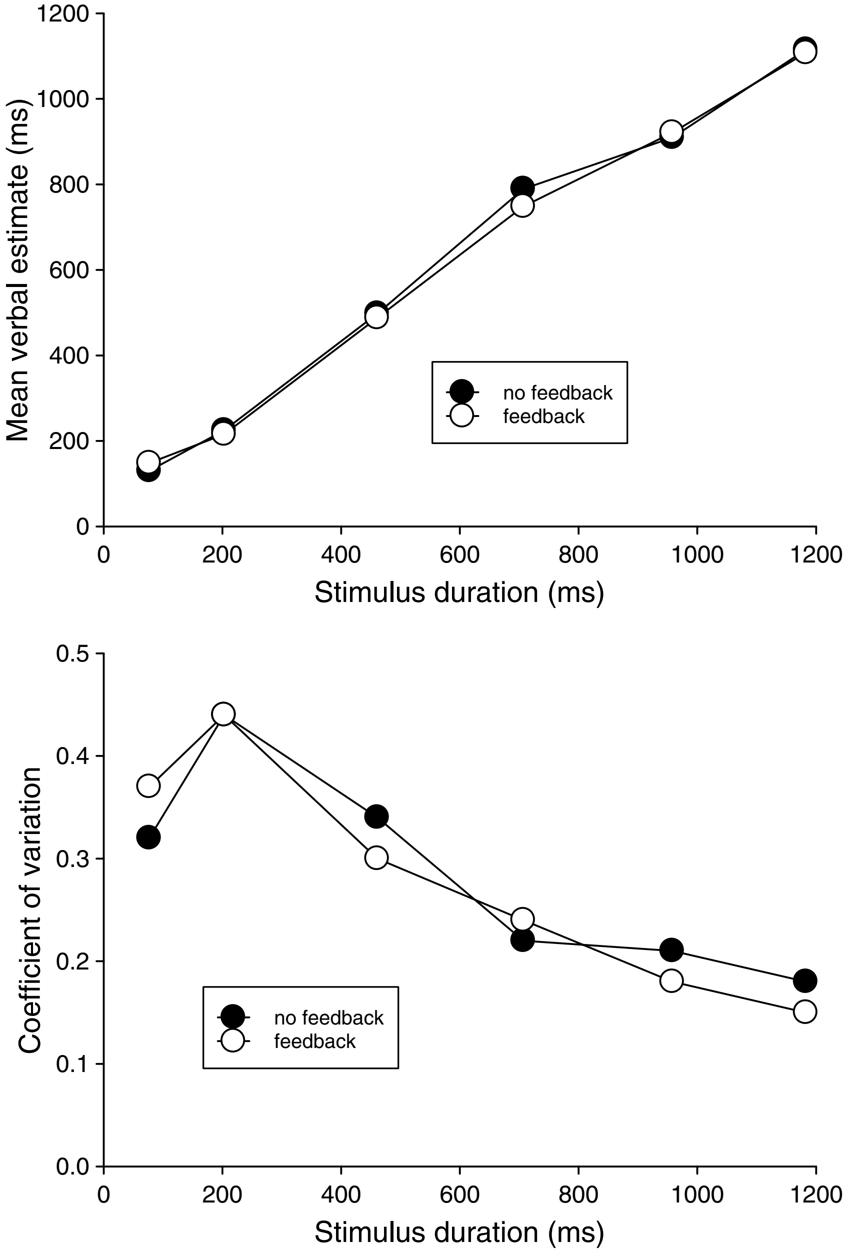

Fig. 1. Upper panel: Mean verbal estimates (ms) plotted against stimulus duration for the feedback (open circles) and no feedback (filled circles) conditions of Experiment 1. Lower panel: Coefficient of variation of duration estimates from Experiment 1 in the feedback (open circles) and no feedback (filled circles) conditions.

without feedback (77–1183 ms). Inspection of data in the upper panel strongly suggests that the NF and F conditions did not produce different means, although participants' estimates changed markedly as a function of stimulus duration. These suggestions were confirmed by ANOVA which found no effect of condition (F versus NF), $F(1,24) = 0.06$, a significant effect of stimulus duration, $F(5,120) = 730.6$, $p < 0.001$, but no F/NF × stimulus duration interaction, $F(5,120) = 0.78$.

The lower panel shows coefficients of variation. Once again, inspection of the data suggests no F/NF effect, but a strong effect of stimulus duration, in this case a general decline in the coefficient of variation from the second-shortest stimulus duration to the longest. These suggestions were confirmed by ANOVA which found no F/NF difference,

$F(1,24) = 0.01$, a significant effect of stimulus duration, $F(5,120) = 13.88$, $p < 0.001$, and no significant F/NF × stimulus duration interaction, $F(5,120) = 0.50$.

The feedback condition had no significant effect upon either mean estimates or coefficients of variation of estimates made by participants. However feedback effects may change behaviour in ways that do not alter the measures shown in Fig. 1. Fig. 2 shows two of these. The upper panel of Fig. 2 shows the average absolute deviation between estimates and the real-time "target" values of the stimuli presented. This was calculated as follows. The absolute difference between the mean estimate produced by a participant at some particular stimulus value and the actual stimulus duration was calculated, and the results averaged together. Inspection of the data in the upper panel of Fig. 2 strongly suggests that

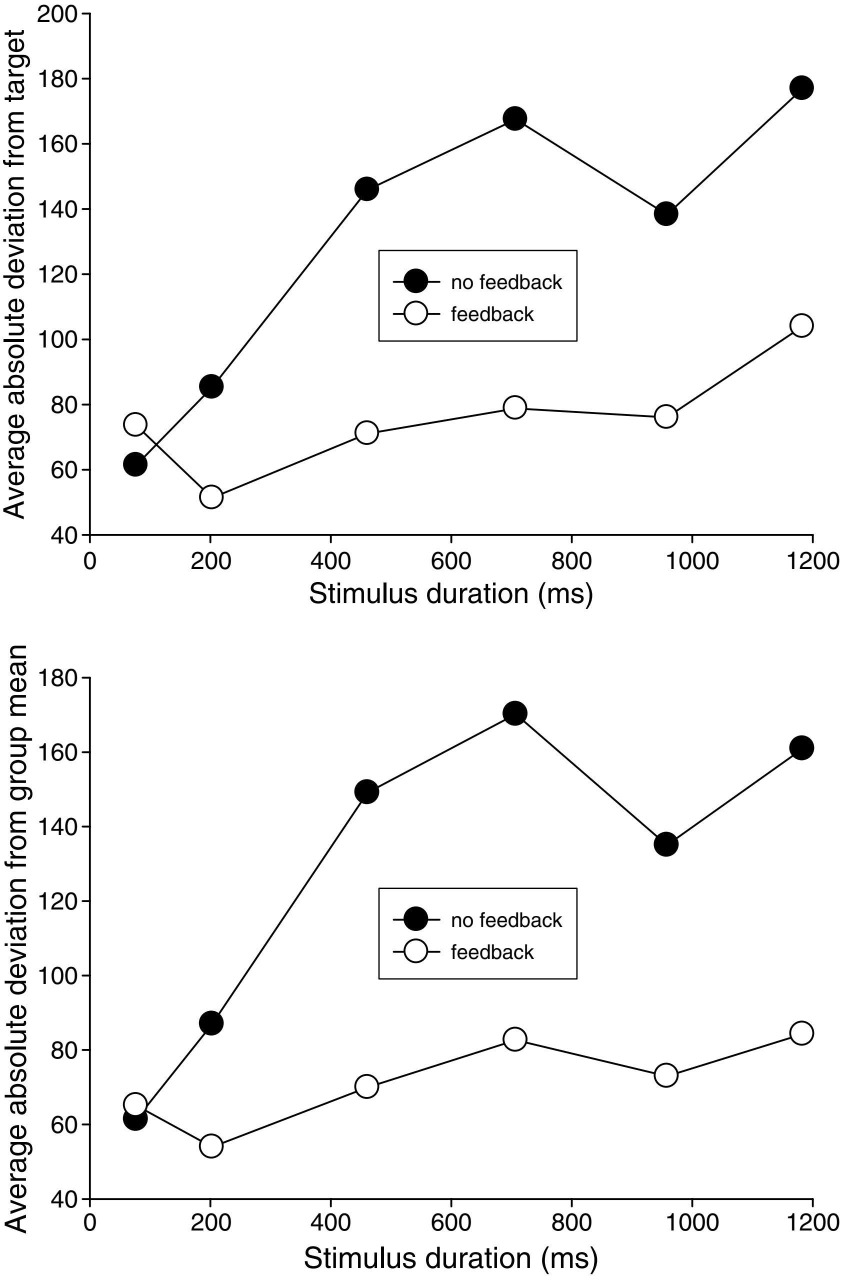

Fig. 2. Upper panel: Mean absolute deviation from target time (actual stimulus duration) from Experiment 1, from the no feedback and feedback conditions. Lower panel: Mean absolute deviation from group mean measure from Experiment 1. See text for the details of calculations.

the provision of feedback reduced the measure, essentially making estimates closer to the target value than without feedback. In addition, the absolute deviation measure appeared to grow with increasing stimulus duration. These suggestions were verified by ANOVA which found a significant effect of feedback, $F(1,24) = 18.31$, $p < 0.001$, as well as significant effects of stimulus duration, $F(5,120) = 5.44$, $p < 0.01$ and a significant feedback/no feedback × stimulus duration interaction, $F(5,120) = 3.96$, $p < 0.01$.

The lower panel of Fig. 2 shows another absolute deviation measure, this time the average absolute deviation between a participant's mean estimate at each stimulus duration value and the group mean, essentially a measure of how similar individuals' mean estimates are to one another. The data bear a striking resemblance to the absolute deviation from target time data shown in the upper panel of Fig. 2, but are actually different. Inspection of the data suggests an F/NF effect, with the feedback condition generally making deviation from the group mean smaller, as well as an effect of stimulus duration, and this was confirmed by ANOVA, which found significant effects of F versus NF, $F(1,24) = 20.69$, $p < 0.001$, and stimulus duration, $F(5,120) = 6.12$, $p < 0.001$, as well as a significant F/NF versus stimulus duration interaction, $F(5,120) = 4.25$, $p < 0.01$.

### 2.3. Discussion

Data from the feedback and no feedback conditions resemble those from previous studies using verbal estimation. The average estimate of the group was highly sensitive to the actual duration of the stimuli presented (cf. Penton-Voak et al., 1996; Wearden et al., 1998, 2006), and average coefficients of variation declined with increasing stimulus duration from the second-shortest duration to the longest (cf. Wearden, 1999, 2003). This suggests that the data collected in Experiment 1 are typical of those obtained using the verbal estimation method to produce non-counting based duration estimates.

A perhaps surprising feature of the results was that the provision of feedback had no overall effect on mean estimates from the group, nor did variability of performance, as assessed by coefficient of variation, change with the provision of feedback. These results contrast with the effects of feedback on mean and variability of time judgements found by Franssen and Vandierendonck (2002). A potential conclusion based on these results is that feedback had no effect on verbal estimation performance, but this is contradicted by the data in Fig. 2 which suggest that feedback had the effect of making each individual participant's mean estimates closer to both the target time (an effect similar to the reduction in "timing error" produced by feedback in some conditions in Franssen & Vandierendonck's study), and to the group mean. On average, therefore, although indices of performance, indicative of accuracy of estimates (mean) and variability of estimates (coefficient of variation) overall from the group were not affected by feedback, the feedback had the effect of making the mean estimates of individuals both closer to the target, and more similar to one another. We will defer a discussion of the processes that might produce these effects until later.

## 3. Experiment 2

Providing feedback on some trials, as in Experiment 1, is only one possible method of "calibrating" participants in verbal estimation tasks, another is to provide specific "anchor" stimuli, whose duration is known, to potentially aid estimates of the duration of stimuli which have to be judged. This was the method used in Experiment 2. As for Experiment 1 there were two experimental conditions, one with "calibrating" stimuli, the other without. On each trial, the participant received two tones, and always had to estimate the duration of the second one (the durations ranged from 77 to 1183 ms, as in Experiment 1), but never the first one. In the "calibration" (C) session, the first stimulus presented took one of the following 5 durations, 50, 200, 500, 800, and 1200 ms, and the participant was informed what the duration was. In the "no calibration" session (NC), the same stimuli were presented, but no information was given as to the duration of the first one.

## 3.1. Method

### 3.1.1. Participants

Fourteen Manchester University Psychology Department undergraduates served for course credit.

### 3.1.2. Apparatus

As Experiment 1.

### 3.1.3. Procedure

Each participant received two experimental sessions. The first one was always the "no calibration" (NC) session, the second the "calibration" (C) session. The procedure for the calibration session was as follows. On each trial the participant received two 500-Hz tones separated by a gap which was a random value between 500 and 1000 ms. The participant was required to estimate the duration of the second tone, using a scale where 1000 = 1 s, and had been informed that all values were between 50 and 1500 ms. A previous instruction read…" you will receive two tones, separated by a gap. You will be told how long the FIRST tone lasted, and the information given to you is always correct. You do not need to estimate the length of the FIRST tone". The second tone (the duration of which had to be estimated) took one of 6 values, 77, 203, 461, 767, 958, and 1183 ms, and the first tone took one of 5 values, 50, 200, 500, 800 and 1200 ms. Before each trial the duration of the first tone was displayed for 500 ms (e.g. "800 ms"). Offset of this display was followed by a random gap ranging from 500 to 1000 ms, then presentation of the first tone of the trial. The participant produced the trials in response to a "Press Spacebar for next trial" prompt, and verbal estimates were typed in, as in Experiment 1. The participant received three blocks of 30 trials (the 6 durations to be estimated × the 5 "calibrating" durations), and 3 blocks were given in all. The order of trials was randomized between blocks with and between participants.

The "no calibration" session was identical except for the omission of the display giving the duration of the first stimulus, and other attendant changes in instructions.

## 3.2. Results and discussion

Data were taken from blocks 2 and 3 of the experiment for both conditions. The upper panel of Fig. 3 shows mean estimates for the group. Inspection of the data suggests little or no effect of calibration session (C/NC) but a strong effect of actual stimulus duration. These suggestions were confirmed by ANOVA which found no effect of C versus NC,

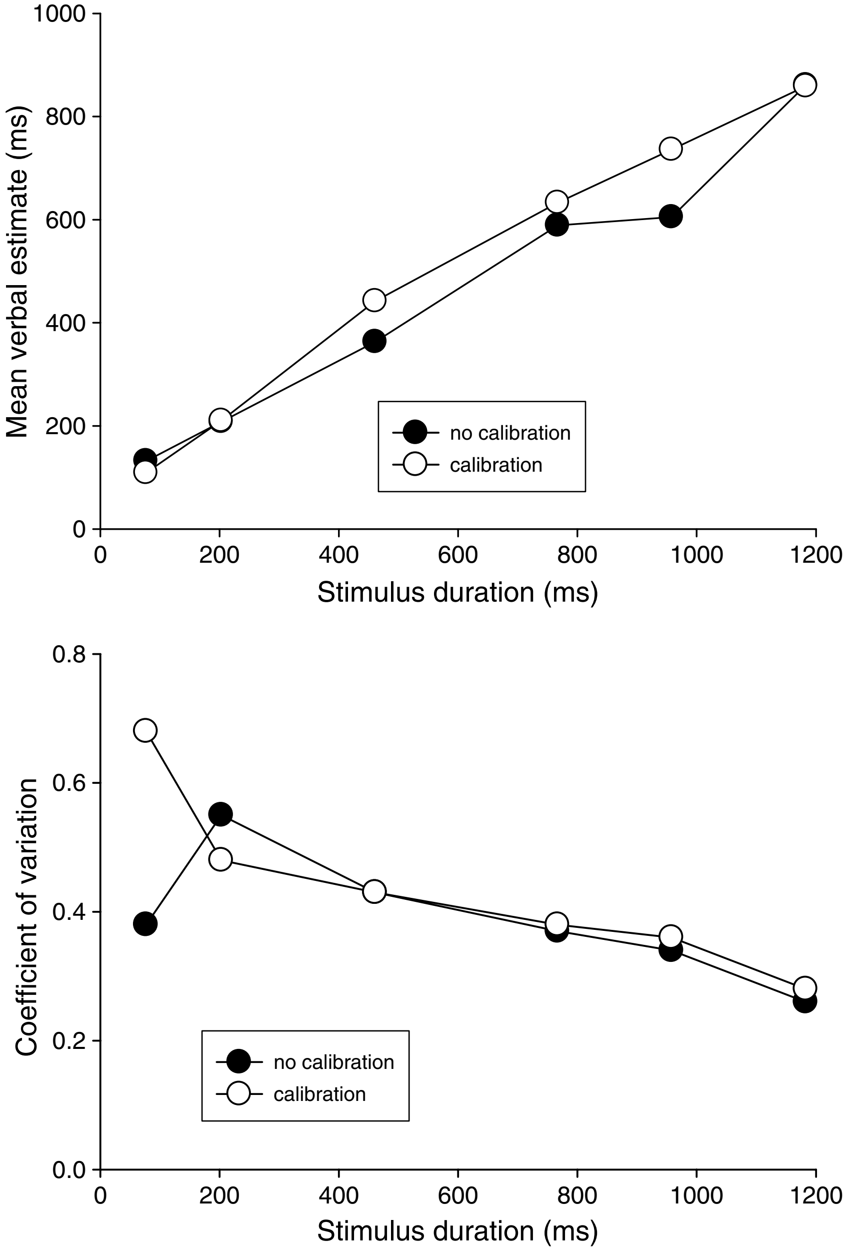

Fig. 3. Upper panel: Mean verbal estimates (ms) plotted against stimulus duration for the calibration and no calibration conditions of Experiment 2. Lower panel: Coefficient of variation of duration estimates from Experiment 2. In both panels, data from the calibration condition are shown as open circles, data from the no calibration condition as filled circles.

$F(1,13) = 0.589$, a significant effect of stimulus duration, $F(5,65) = 191.42$, $p < 0.001$, but no C/NC × stimulus duration interaction, $F(5,65) = 1.43$. The lower panel of Fig. 3 shows coefficients of variation of estimates. Once again, inspection suggests little or no effect of calibration, but an effect of stimulus duration, and this was confirmed by ANOVA. There was no effect of C versus NC, $F(1,13) = 1.33$, but a significant effect of stimulus duration, $F(5,65) = 5.87$, $p < 0.001$. The interaction between calibration condition and stimulus duration was not significant, but approached significance, $F(5,65) = 2.05$, $p = 0.08$.

The upper panel of Fig. 4 shows the absolute deviation from target time measure used in Experiment 1. Inspection of the data suggests that the calibration manipulation reduced the deviation between estimates and the target time, and that this deviation grew

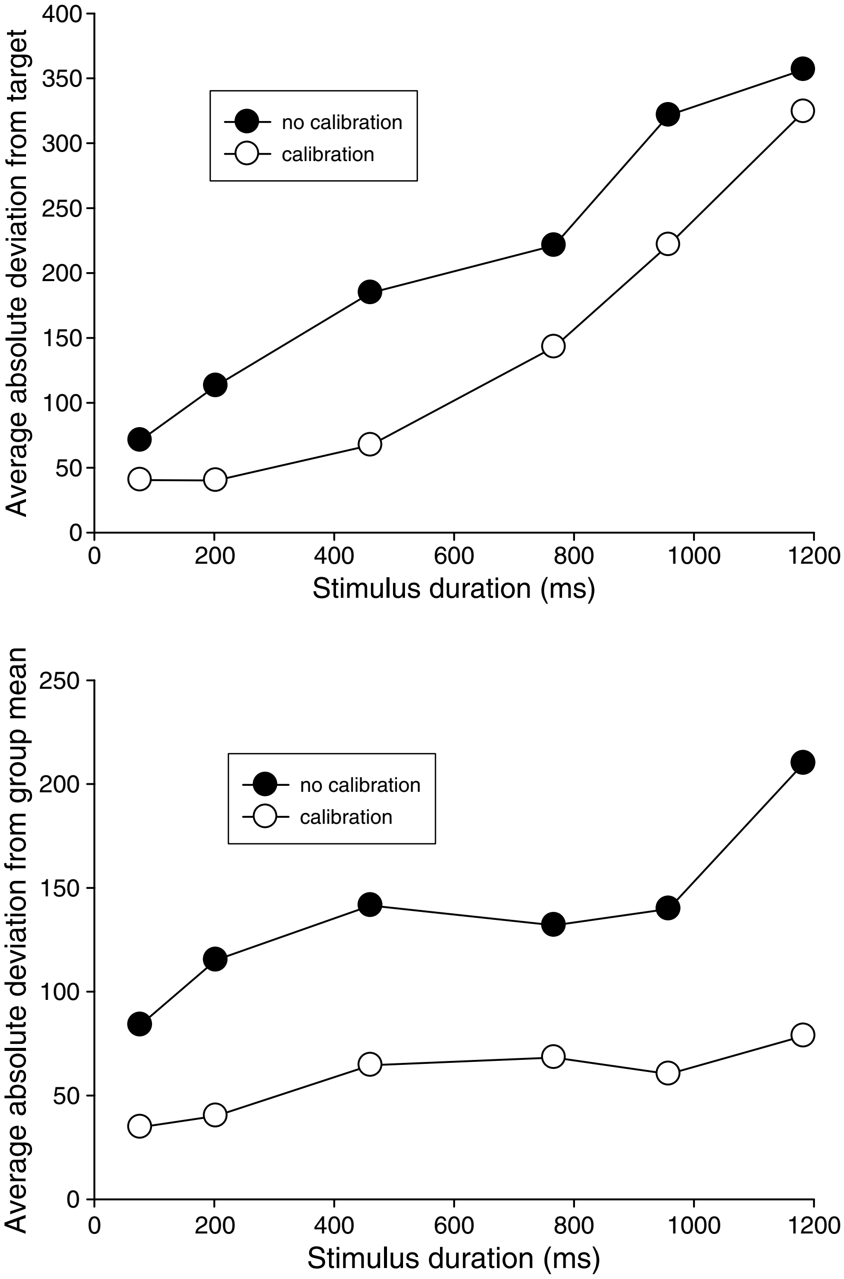

Fig. 4. Upper panel: Mean absolute deviation from target time (actual stimulus duration) from Experiment 2, from the calibration and no calibration conditions. Lower panel: Mean absolute deviation from group mean measure from Experiment 1. See text for the details of calculations.

with stimulus duration. Both these suggestions were confirmed by ANOVA which found significant effects of calibration condition (NC/C), $F(1,13) = 8.55$, $p < 0.05$, and stimulus duration, $F(5,65) = 32.35$, $p < 0.001$, but no NC/C × stimulus duration interaction, $F(5,65) = 1.10$.

The lower panel of Fig. 4 shows the deviation from the group mean measure used in Experiment 1. Calibration appeared to make deviation from the group mean much smaller than in the no calibration session, and this was confirmed by ANOVA. There were significant effects of calibration, $F(1,13) = 10.98$, $p < 0.01$, and stimulus duration, $F(5,65) = 5.52$, $p < 0.001$, but no significant calibration × stimulus duration interaction, $F(5,65) = 1.33$.

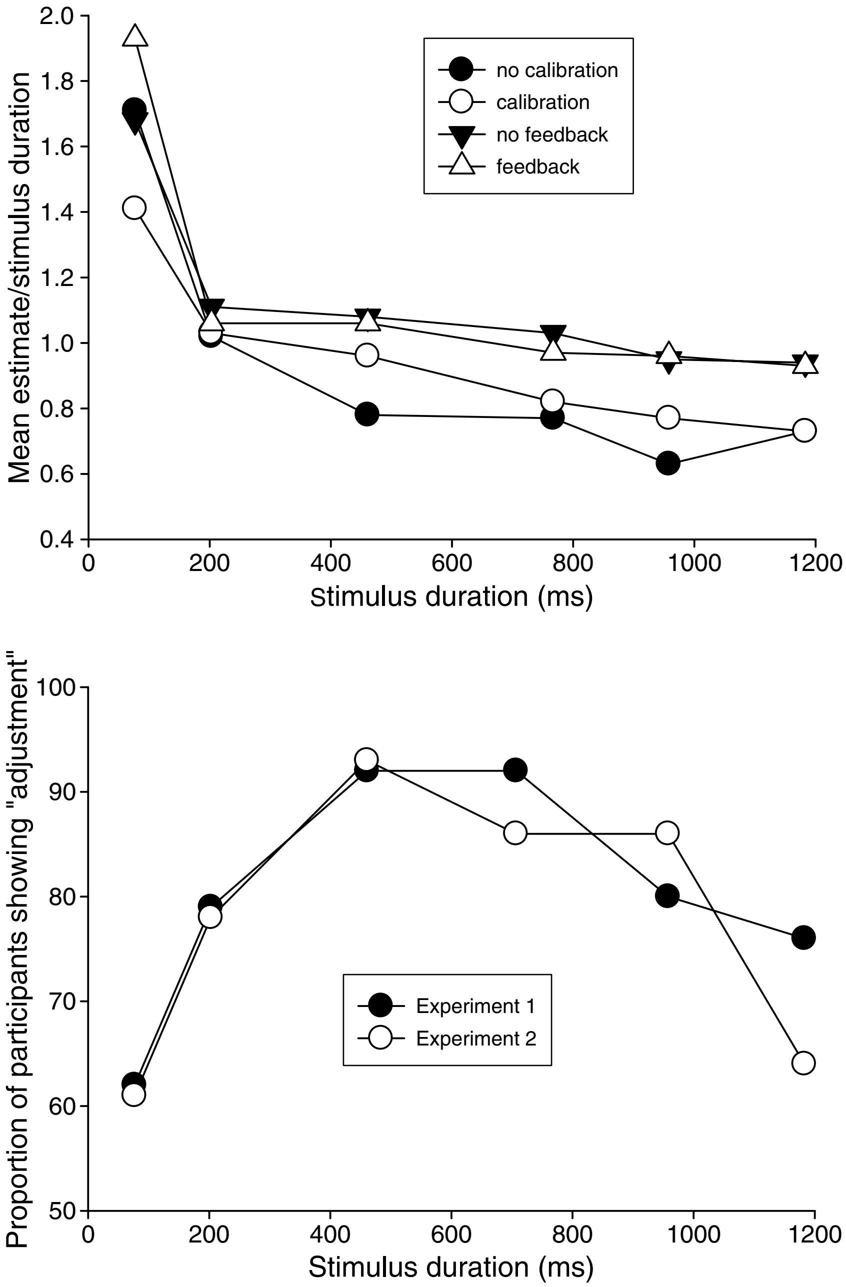

Fig. 5. Upper panel: Mean verbal estimates from all conditions of Experiments 1 and 2 divided by actual stimulus duration. Values greater than 1.0 show overestimation, values less than 1.0 underestimation. Lower panel: proportion of participants showing "adjustment" in Experiments 1 and 2. See text for details of calculation. In both panels, measures are plotted against actual stimulus duration.

### 3.3. Between-experiment analyses

Some effects in the data from Experiments 1 and 2 emerge most clearly when data from the two studies are presented together, and one of these is shown in the upper panel of Fig. 5.

To produce the results shown mean verbal estimates from each condition were divided by the real-time value of the duration estimated. Now, values greater than 1.0 show overestimation, a value of 1.0 shows accurate estimation, and values below 1.0 show underestimation. In all the conditions of Experiments 1 and 2, the data on average conform to

Vierordt's Law, the proposition that short durations are overestimated and longer ones underestimated. The "indifference point" (i.e. accurate estimation) comes between about 300 and 700 ms for the different conditions. Conformity to Vierordt's Law has been found to be typical of data collected with a verbal estimation method in this duration range, see Wearden (2003).

Another issue which can be clarified by considering the data from the two experiments together is how the effect of feedback in both can be captured quantitatively. The lower panel of Fig. 5 illustrates the results of a potential method of description. For each participant at each stimulus duration in Experiments 1 and 2, we first asked whether the mean estimate was above or below the actual stimulus duration in the NF (Experiment 1) or NC (condition), that is, *without* the putative calibrating operation. Then we examined the mean from the same stimulus duration value in the F and C conditions, *with* the putative calibrating operation. The measure used was "adjustment". "Adjustment" was defined as a *shorter* estimate in the F or C condition if the estimate in the NF and NC conditions had been *above* the target time, or a *longer* estimate in the F and C condition if the estimate in the NF and NC conditions had been *below* the target time. The other possible sort of change was an increase if the NF and NC estimates had been above the target time, and a decrease if the NF and NC estimates had been below it. In a few cases, which were ignored, the estimates from the uncalibrated and calibrated conditions were exactly the same.

The lower panel of Fig. 5 shows the proportion of participants showing "adjustment" at each target time. Obviously, there was a very strong effect for "too high" estimates to be reduced by feedback/calibration, and "too low" estimates to be increased by it, particularly at the intermediate stimulus durations. The overall percentage of "adjustment" was 83% for Experiment 1 and 82% for Experiment 2, showing that the performance of participants was very strongly influenced by the calibration operations in spite of the lack of effect on means and coefficients of variation shown in Figs. 1 and 3.

The fact that "adjustment" very often occurred does not mean that participants were invariably closer to the target time in the feedback and calibration conditions than without this manipulation, although there was overall a tendency for this to happen, as the upper panels of Figs. 2 and 4 show, which was highly significant statistically. For example, for the 767 ms stimulus duration, a person could make an estimate of 800 ms without feedback/calibration and an estimate of 700 ms with it. This would count as an "adjustment", as the estimate was above the target then decreased, but the estimate with feedback/calibration is actually *further* from the target time than without it. Likewise, the existence of "adjustment" does not necessarily mean that the slope of the function relating mean estimate to stimulus duration was always closer to 1.0 with feedback/calibration than without it. For example, we calculated the regression of mean estimate versus stimulus duration for all the participants in Experiment 2 from the NC and C conditions, and found that regression slopes in the C condition were closer to 1.0 than in the NC condition in only 8 of the 14 cases. Fig. 6 shows data from two individual participants illustrating the variability of the effect of the calibration manipulation.

The upper panel of Fig. 6 shows data from the participant who was by far the most insensitive to stimulus duration in the NC condition of Experiment 2, with estimates only slightly more than doubling when stimulus durations from 77 ms to 1183 ms were presented. The slope of the regression line of mean estimate versus stimulus duration was only 0.2. The calibration clearly made the participant's estimates much more sensitive to stimulus duration as well as more accurate: in the C condition, the estimate of the 1183 ms

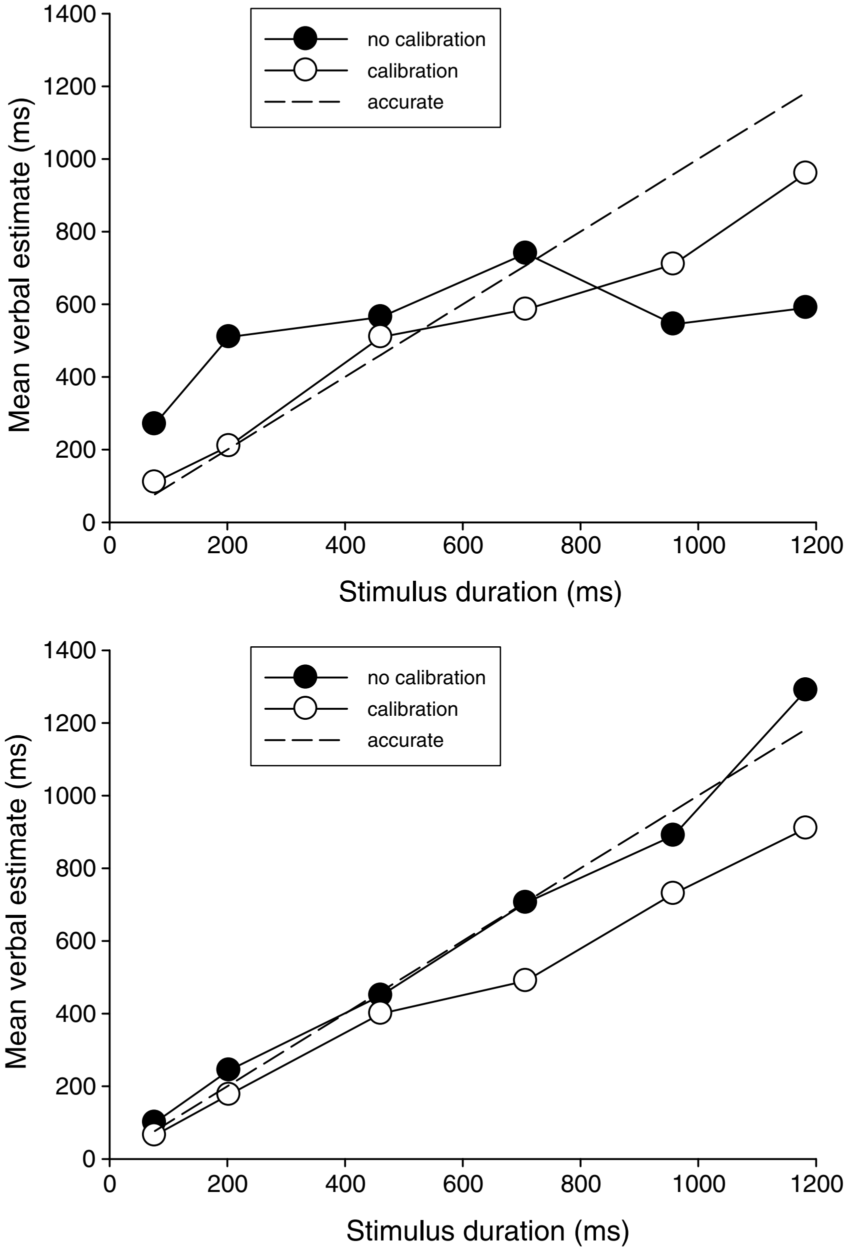

Fig. 6. Data from two participants in Experiment 2. Data from the calibration and no calibration conditions are shown separately. The diagonal dotted line shows accurate estimation.

stimulus was nearly nine times as great as the estimate of the 77 ms one, as well as all estimates but one being closer to the target time in the C condition than the NC one, and the slope of the regression line accordingly increased to 0.71. The lower panel, however, shows data from a participant who was highly accurate in the NC condition (regression slope = 1.00). Now, the calibration made all estimates but one *more* deviant from the target time than in the NC condition, as well as reducing sensitivity to stimulus duration (the regression slope in the C condition was 0.73).

## 4. General discussion

The results from Experiments 1 and 2 can be simply summarized. Firstly, mean verbal estimates from the groups of participants overestimated actual stimulus duration when this was short, then were shorter than or close to actual stimulus value when this was longer, showing conformity to Vierordt's Law in both experiments (Fig. 5), and coefficients of variation significantly declined with increases in stimulus duration (Figs. 1 and 3). In both these respects, data from our verbal estimation studies violate the mean accuracy and scalar variance properties required by SET, but this violation is in fact completely typical of data collected with the verbal estimation method in this stimulus range (Wearden, 1999, 2003). Neither feedback (Fig. 1) nor the calibration operation (Fig. 3) had any significant effect on mean verbal estimate or average coefficient of variation. On the other hand, both feedback (Fig. 2) and calibration (Fig. 4) made the mean verbal estimates from individuals (a) closer to the actual duration presented and (b) closer to the group mean (i.e. more similar). Both feedback and calibration resulted in a process of "adjustment" where individual participant's estimates that were above or below the actual stimulus duration in the no feedback or no calibration condition tended to increase or decrease in the appropriate direction (i.e. decrease if above, increase if below) with feedback or calibration (Fig. 5), although this did not necessarily mean that an individual's estimates would be more accurate with feedback/calibration than without (e.g. see Fig. 6).

The results in some ways seem paradoxical: how can feedback/calibration change some measures apparently quite radically, while leaving conventional performance measures like mean and coefficient of variation unchanged? Fig. 7 shows one way in which this is possible.

To develop the argument, we simplify both the task and the invented participants' performance. Assume that the durations to be estimated are 300, 500, 700, 1000, and 1200 ms. Imaginary results are shown from two participants who constitute a "mini-group" of two. In the *without* condition (without feedback or calibration) one of the participants (above without) overestimates all durations, and the other (below without) underestimates all of them. In the *with* conditions (with feedback or calibration), the "above" participant reduces all estimates, and the "below" participant increases them all. In this case, the "mini-group" mean estimate is completely unaffected by the presence of feedback or

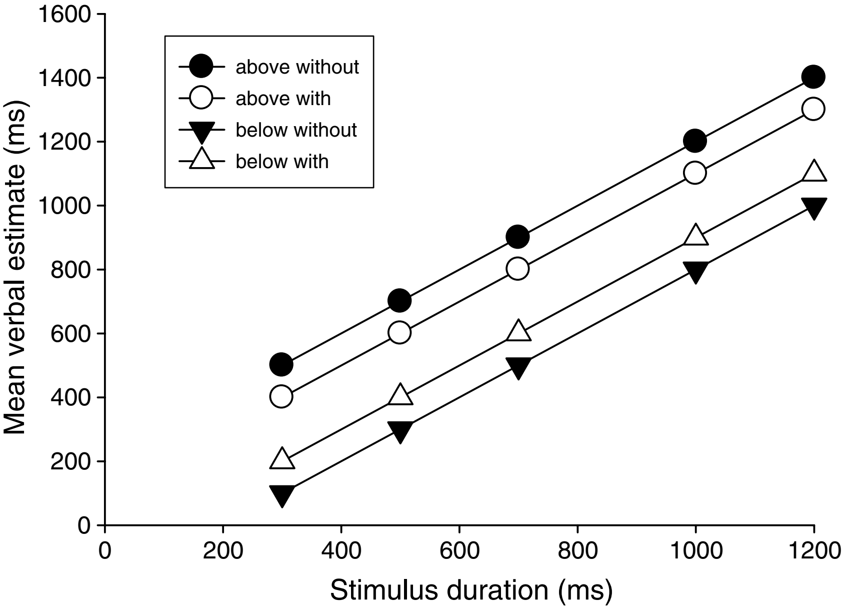

Fig. 7. "Proto-model" showing adjustment of verbal estimate values with feedback/calibration without change in group mean. See text for details.

calibration (as one participant increases estimates and the other decreases them), but the deviation of individual participant means from the target time, and deviation from the group mean, is reduced, exactly the effect obtained in data. The "proto-model" shown in Fig. 7 may also account for the lack of change of coefficient of variation in the different conditions. Data in the lower panels of Figs. 1 and 3 show that coefficients of variation decline with increasing stimulus duration, but the decline locally may be small. For example, if a participant changes his/her estimate of some duration, *t*, from 100 ms to 1000 ms, in different conditions, then a change in coefficient of variation would be expected to be observed, but not if they change their estimate from 500 to 600 ms. In other words, small and possibly realistic changes in estimates occasioned by feedback or calibration may produce little or no change in the observed coefficients of variation.

However, although the invented results in Fig. 7 can reconcile processes of adjustment with absence of change in overall group mean and coefficient of variation, they are only illustrative, and are not generally representative of the behaviour of individuals, who rarely consistently either under- or over-estimate real stimulus durations at all time values (see Fig. 6). Modelling the specific details of the way that people use feedback/calibration in the present investigation is hampered by the fact that we have no model of how verbal estimation of duration is carried out in the first place. Consider the effect of feedback (Experiment 1). Unlike in the case of interval production (cf. Franssen & Vandierendonck, 2002; Wearden et al., 1997, Experiment 4), feedback did not change group mean estimates and, in general, it seems likely that feedback had more complex effects on verbal estimation than in interval production, where a participant can receive a series of identical intervals to produce and experience feedback after each. Such a procedure seems likely to give rise to a change in the participant's "standard" for the task, or some form of sequential adjustment from trial to trial. However, even in the case of interval production the effects of feedback are not at present fully understood (Franssen & Vandierendonck, 2002).

In the case of feedback after some verbal estimates, as in Experiment 1, the participant cannot simply use the feedback after one trial to control performance on the next, as the stimulus durations to be judged may differ radically from one trial to the next. Furthermore, the "Vierordt" tendency noted in some verbal estimation data (Fig. 5, see also Wearden, 2003) means that feedback may not always guide estimates in the direction of greater accuracy. A person is on average likely to overestimate a short duration, and if feedback after this sort of trial produces a tendency to reduce estimates in general, this may shorten estimates of longer durations, taking them further away from accuracy than before. The converse effect may also occur. Added to this complex potential relation between performance and the effect of feedback is then the tendency of participants to use only certain "quantized" values (usually ending in "00") when making estimates of stimuli in the duration range employed in the present study. This means that feedback may not exert a progressive refinement of performance (as in interval production where productions that are too short or long can be progressively changed over trials) but "jumps" from one "quantized" estimate to another.

Overall, therefore, any attempt to model the data obtained in the present study in any kind of reasonable detail seems both exceptionally difficult, and many questions remained unanswered. For example, mean verbal estimates from individuals frequently grow as a linear function of the real duration estimated, but slopes and intercepts of such linear growth are often less than 1.0, and greater than zero, respectively. Does this deviation from the linear and accurate "estimation" assumed by SET depend on some sort of non-veridical "scale" used by the participant, or does it result from the way a basically linear and on average accurate scale is "quantized"? For example, even if the lowest duration value used in the present study (77 ms) was represented on average accurately in terms of some internal scale, it might be "overestimated" if the smallest estimate that a participant would actually make is 100 or 200 ms. Likewise, failures to use values at the upper end of the permissible range (50–1500 ms) may result in the "underestimation" of longer durations, even though these may be scaled "correctly" in terms of some internal scale of "temporal sensation". On the other hand, it may be that the scale participants are using is not on average accurate, and that the quantization contributes mainly to the pattern of variance observed. Future computer modelling may help decide among the many possibilities that exist for scaling and quantization operations in verbal estimation, and understanding of basic mechanisms may then permit the development of a proper quantitative model of how feedback and calibration change verbal estimates.

## References

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

Allan, L., & Gibbon, J. (1991). Human bisection at the geometric mean. *Learning and Motivation, 22*, 39–58.

Droit-Volet, S., & Izaute, M. (2005). The effect of feedback on timing in children and adults: The temporal generalization task. *Quarterly Journal of Experimental Psychology, 58A*, 507–520.

Fraisse, P. (1964). *The psychology of time*. London: Eyre and Spottiswoode.

Franssen, V., & Vandierendonck, A. (2002). Time estimation: does the reference memory mediate the effects of knowledge of results? *Acta Psychologica, 109*, 239–267.

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., Meilleur-Wells, G., & Lachance, R. (1999). When to start explicit counting in a time-interval discrimination task. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 993–1004.

Penton-Voak, E. P., 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*, 307–320.

Wearden, J. H. (1991). Human performance on an analogue of an interval bisection task. *Quarterly Journal of Experimental Psychology, 43b*, 59–81.

Wearden, J. H. (1992). Temporal generalization in humans. *Journal of Experimental Psychology: Animal Behavior Processes, 18*, 134–144.

Wearden, J. H. (1999). Beyond the fields we know…: exploring and developing scalar timing theory. *Behavioural Processes, 45*, 3–21.

Wearden, J. H. (2003). Applying the scalar timing model to human time psychology: progress and challenges. In H. Helfrich (Ed.), *Time and Mind II: Information-processing perspectives* (pp. 21–39). Gottingen: Hogrefe & Huber.

Wearden, J. H. (2004). Decision processes in models of timing. *Acta Neurobiologiae Experimentalis, 64*, 303–317.

Wearden, J. H., & Lejeune, H. (submitted for publication). Scalar properties in human timing: Conformity and violations. *Quarterly Journal of Experimental Psychology*.

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, 51B*, 97–120.

Wearden, J. H., & McShane, B. (1988). Interval production as an analogue of the peak procedure: evidence for similarity of human and animal timing. *Quarterly Journal of Experimental Psychology, 40B*, 363–375.

Wearden, J. H., Todd, N. P. M., & Jones, L. A. (2006). When do auditory/visual differences in duration judgements occur? *Quarterly Journal of Experimental Psychology, 59*, 1709–1724.

Wearden, J. H., Wearden, A. J., & Rabbitt, P. (1997). Age and IQ effects on stimulus and response timing. *Journal of Experimental Psychology: Human Perception and performance, 23*, 962–979.