ψ Psychology Press
Taylor & Francis Group

Scalar properties in human timing: Conformity and violations

J. H. Wearden
Keele University, Keele, UK

Helga Lejeune
University of Liège, Liège, Belgium

Data from studies of timing in human participants were reviewed with respect to their conformity to the two scalar properties of timing: mean accuracy and the scalar property of variance. Results reviewed were taken from studies of temporal generalization, temporal bisection, discrimination methods, and "classical" timing tasks such as the reproduction, production, and verbal estimation of duration. Evidence for one or both scalar properties was found in many studies, including those using children and elderly participants, but systematic violations were sometimes noted. These violations occurred (a) when very short durations (<100 ms) were timed, (b) in situations in which timing tasks varying in difficulty were compared, (c) when classical timing tasks were employed, and (d) in situations where highly practised observers exhibited unusual patterns of variance. A later section attempted to reconcile some of these violations with an underlying scalar-consistent timing mechanism.

The present article reviews results from studies with human participants with respect to their conformity to, or violation of, the *scalar properties* of timing. A companion article (Lejeune & Wearden, 2006) performed a similar function for results from experiments with nonhuman animals (hereafter just "animals"). The review of material from experiments using humans or animals was separated mainly because the techniques involved, and some of the theoretical issues, differ in animal and human research. Although research initially conducted with animals has inspired a large body of recent work on human timing (see Allan,

1998, and Wearden, 2003, for reviews), the inspiration has been mostly theoretical, concerned with the application of ideas derived from Gibbon, Church, and Meck's (1984) *scalar expectancy theory* (SET). However, parallels between effects noted in research with humans and animals are interesting, and a later section discusses some of these.

Lejeune and Wearden (2006) made the distinction between *empirical scalar timing*, *theoretical scalar timing*, and *predictions of SET*, and these distinctions also hold here. We first discuss how the scalar properties of timing are defined.

1

Scalar properties of timing: Definitions and measurement

Behaviour that conforms to the scalar properties of timing exhibits two properties. The first is *mean accuracy*, the requirement that mean measures of timed behaviour vary linearly, and usually accurately, with imposed time requirements or temporal standards. So, if a person produces time intervals of some specified value, mean accuracy would require that the times produced on average equal the time requirement. If people are judging the duration of presented stimuli they should behave as if their judgements are on average accurate reflections of real stimulus duration. In some cases, however, the task may require that the participants make a timed response at some constant fraction of the standard or target duration to be timed, in which case mean accuracy requires a simple proportional relation between response measures and the real-time values of the durations used.

The second scalar property is the *scalar property of variance*, a form of Weber's law, which requires timing sensitivity to remain constant as durations timed vary. There are three main ways of testing this property. The simplest is to take the standard deviation (s) and the mean (m) of behaviour measures when some time constraint or requirement, t, is varied. Regression of s against m should yield a linear function with a high r^2, to conform to the scalar property. Another method of testing the scalar property is to use the s and m from each duration timed to construct a *coefficient of variation* ($CV = s/m$), a Weber-fraction-like measure. Now, the scalar property of variance requires CV to remain constant as t is varied. A technical problem with this method is that it tests the scalar property of variance essentially by obtaining a *negative* result (e.g., a nonsignificant slope in regression of CV against t, or nonsignificant differences between the CV values obtained with different t), which some readers may find problematical. In spite of this difficulty, plotting CV against t can produce some striking examples of conformity to the scalar property, as in Wearden's (1991a, p. 64) plots of data derived

from Zeiler, Scott, and Hoyert (1987), which found CV values close to 0.15 for t values from 0.5 s to 32 s. Some studies also employ other Weber-fraction-like measures, and, as for coefficients of variation, conformity to the scalar property requires that these measures remain constant as the absolute durations timed vary between conditions.

The final commonly used method is tests of *superimposition* (called *superposition* in the U.S. literature), which involves plotting measures of timed behaviour from judgements of different absolute times on the same *relative* scale. The appropriate form of the relative scale depends on the task used, but a simple example comes from Wearden's (1992) study of temporal generalization in humans. Here, people are initially presented with (or have to initially learn) a temporal standard, for example a tone 400 ms long. After initial standard presentations, people then receive comparison stimuli (e.g., tones from 100 to 700 ms in duration in 100-ms steps) and have to judge whether each comparison is the standard (making a YES or NO response). The proportion of YES responses plotted against stimulus duration constitutes a temporal generalization gradient.

Figure 1 presents invented (but very typical) results to illustrate the normal shape of temporal generalization gradients obtained from humans. In the upper panel three conditions are shown. Two of these involve the same relative spacing around the standard duration, so in one the standard is 400 ms, and the comparisons are spaced in 100-ms steps around it (i.e., from 100 to 700 ms), and the other has a standard of 800 ms, with 200-ms comparison steps. Both generalization gradients (a) peak at the standard and (b) are slightly asymmetrical, with more YES responses at comparisons longer than the standard than shorter than it. Both these features are completely typical for temporal generalization in humans (e.g., Wearden, 1992). The 400/100 and 800/200 values were invented so that their temporal generalization gradients would superimpose when plotted on the same relative scale, where each the value of comparison duration is divided
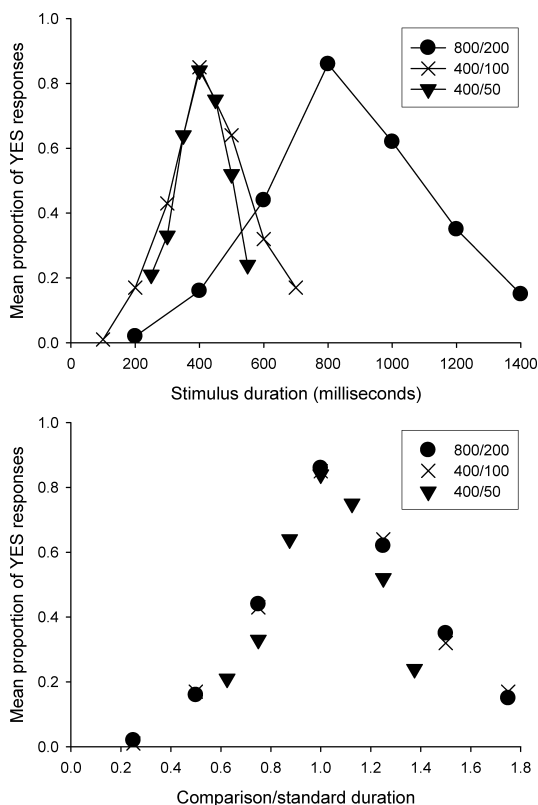
Figure 1. *Illustration of temporal generalization results. Upper panel: temporal generalization gradients in the form of proportion of YES responses (identifications of a stimulus as having the standard duration) plotted against stimulus duration. Results are shown separately for three conditions: 800/200 (800-ms standard, 200-ms comparison spacing); 400/100 (400-ms standard, 100-ms comparison spacing), and 400/50 (400-ms standard, 50-ms comparison spacing). Lower panel: Results from the upper panel plotted on the same relative scale, where each comparison duration is divided by the standard duration in force for the condition (400 or 800 ms).*

by that of the standard in force, and this is shown in the lower panel of Figure 1.

In his Experiment 3, Wearden (1992) used standard values of 500, 600, and 700 ms and found that gradients were superimposed when plotted on a scale where each comparison duration was expressed as a fraction of the standard in force (see his Figure 4, p. 139). The quality of superimposition is very often determined simply by inspection of the plot, but statistical tests can be used

where data from different conditions, plotted on the same relative scale, are compared. As in the case of CV differences, discussed above, superimposition requires a statistically *negative* result from this sort of comparison.

If observed behaviour directly exhibits the two scalar properties, we refer to this as *empirical scalar timing*. However, another possibility is what Lejeune and Wearden (2006) call *theoretical scalar timing*. This is where underlying time representations can be considered to have scalar properties, but the observed behaviour does not, usually as the result of "additional processes" modulating the expression of underlying scalar timing. We are concerned with reconciling behaviour with underlying scalar timing in a later section of the article. The final form of conformity to scalar timing is detailed conformity to SET's predictions of performance on some particular task, but this is not our focus here. The reason for concentrating on empirical and theoretical scalar timing is that these two properties are more general than SET itself.

Competitors to SET, which exist mainly in the domain of animal timing (e.g., Killeen & Fetterman, 1988; Kirkpatrick, 2002; Machado, 1997; Staddon & Higa, 1999), all use some mechanism for generating empirical scalar timing, albeit mechanisms different from those proposed by SET, so a review concentrating on empirical and theoretical scalar timing (i.e., behaviour that directly or indirectly reflects scalar properties of time representations) can identify conformities that all theories need to explain, as well as violations that all theories may need to take into account to be comprehensive. However, although we are not here concerned with the specific predictions of SET in particular cases, the instantiation of SET in terms of cognitive processes (discussed briefly in a later section) can furnish us with some tools that are useful in accounting for what appear at first to be deviations from the scalar properties of time.

An important consideration in work with humans is the possible presence of chronometric counting (e.g., dividing the duration into countable subunits, which are then sequentially labelled

and produced overtly or covertly). Humans will use chronometric counting at durations greater than about 1.2 s (Grondin, Meilleur-Wells, & Lachance, 1999), and modelling of the counting process (Killeen & Weiss, 1987; Wearden, 1991a) shows that although the mean accuracy property will usually hold when counting is used, the scalar property of variance will not: When counting is used, CVs decline systematically with increasing duration (see discussion in Wearden, 1991a, pp. 68–72).

Three methods of avoiding counting have been used in the literature. One is to employ short durations (usually <1 s), another is to prevent counting by a concurrent interfering task, and the last is to employ children of less than about 8 years of age as participants, as children of this age do not spontaneously count when timing. The commonest form of concurrent interference used requires participants to repeat aloud digits briefly presented on a computer screen with temporally irregular spacing (e.g., Wearden, Denovan, Fakhri, & Haworth, 1997a, and Wearden, Rogers, & Thomas, 1997b). This effectively prevents counting, but the irregularity of presentation also makes the number of digits repeated a very unreliable guide to stimulus duration. All three methods have been used in recent work inspired by SET, but many studies from the classical timing literature have not controlled for counting.

The study of the sensitivity of human behaviour to temporal constraints, classically conducted under the heading of *time perception* (e.g., Fraisse, 1964), provides innumerable examples of behaviour that might or might not exhibit the scalar properties of timing. To review more than a fraction of this work would be impossible in a single article, and we concentrate particularly on those studies whose data can be readily discussed in terms of the two scalar properties of timing. In the review below, we discuss separately data coming from four main areas: temporal generalization, bisection, discrimination methods, and "classical" procedure such as verbal estimation, production, and reproduction. In all these cases many studies provide data enabling the scalar properties of timing to be tested.

Temporal generalization

Temporal generalization poses participants what seems to be one of the simplest possible questions about duration, namely whether one duration is equal to another. As discussed above, the normal method used initially presents some temporal standard, then comparison durations equal to the standard, longer, or shorter are presented, with a YES/NO response (standard or not?) measuring behaviour. Generalization gradients (the proportion of YES responses plotted against comparison stimulus duration) can be analysed for both the scalar properties of timing. Mean accuracy is indicated by location of the peak of the temporal generalization gradient at the standard duration, and the scalar property of variance is demonstrated when gradients can be superimposed by being plotted on the same relative scale (i.e., with each comparison duration divided by the standard in force).

As mentioned previously, Wearden (1992) varied the temporal generalization standard over values of 400, 500, 600, and 700 ms and found that the peak of the temporal generalization gradient tracked the standard perfectly, with superimposition also being found (see his Figure 4, p. 139). Likewise, Wearden et al. (1997a) used standards of 2, 4, 6, and 8 s in a study where counting was prevented by a secondary task. The generalization gradients peaked at the standard, and gradient superimposition was almost perfect. Thus, temporal generalization performance with both short (<1 s) and longer (2–8 s) standard values exhibited both mean accuracy and the scalar property of variance in the studies quoted.

Mean accuracy was found in elderly participants in Wearden, Wearden, and Rabbitt (1997c) and McCormack, Brown, Maylor, Darby, and Green (1999), who also used children of 5, 8, and 11 years, all employing a 400-ms standard. Wearden and Towse (1994) provide other examples of mean accuracy, including one condition where the peak of the generalization gradient was at the standard even when the standard was the shortest or longest of all the stimuli presented, thus showing that participants appear to be using the standard as the basis for their decisions

rather than, for example, the mean or midpoint of all the comparison durations presented.

Droit-Volet, Clément, and Wearden (2001) tested children of 3, 5, and 8 years with 4-s and 8-s standards. Gradients peaked at the standard in all cases except the younger children with a 4-s standard, where the peak of responding was at 3 s, and at all ages superimposition was found (their Figure 3, p. 279). Droit-Volet (2002) compared temporal generalization performance in 3- to 8-year-old children with 400-ms and 4-s standards and reported temporal generalization gradient peaks showing close to mean accuracy, and good super-imposition, over the 10-fold range of durations employed (see her Figure 2, p. 1200).

However, superimposition is absent in some sorts of comparisons of gradients produced by auditory and visual stimuli, with auditory gradients being steeper than the visual gradients, even when durations were the same, indicating failure of superimposition, see Wearden, Edwards, Fakhri, and Percival (1998, their Figure 1, p. 102). Another failure of superimposition comes from Ferrara, Lejeune, and Wearden (1997). People performed a temporal generalization task with a 600-ms standard. In one group, nonstandard dur-ations were spaced in 150-ms steps ("easy") around the standard, and in another group they were spaced in 75-ms steps ("difficult"). The generaliz-ation gradients from the "easy" and "difficult" con-ditions did not superimpose (their Figure 1, p. 222), with more sensitive timing (i.e., steeper temporal generalization gradients) being exhibited in the "difficult" condition. Although the superim-position property was violated, mean accuracy held, as peaks of the generalization gradients were found at the standard duration (see also Wearden & Grindrod, 2003). The case found in Ferrara et al. (1997) and Wearden and Grindrod (2003) is illustrated in Figure 1 with invented results coming from a condition (400/ 50), in which the standard duration is 400 ms, and comparisons are spaced in 50-ms steps around it. Now, the generalization gradients are steeper than those with 100-ms steps, as shown in both panels of Figure 1, so superimposition is violated.
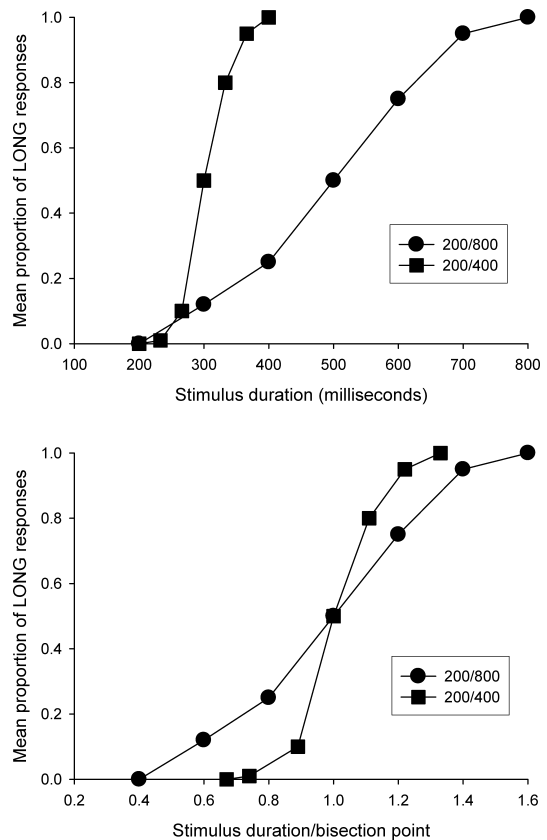

Figure 2. *Illustration of bisection results. Upper panel: Psychophysical functions (proportion of LONG responses plotted against stimulus duration) for two bisection conditions, one with a 200/800 short/long (SL) pair, the other with a 200/400 SL pair. Lower panel: Superimposition plot for the two psychophysical functions shown in the upper panel. Each stimulus duration is divided by the bisection point for the condition shown (500 ms for 200/800, 300 ms for 200/400).*

A variant of the normal temporal generalization procedure was developed by Wearden and Bray (2001), their "episodic temporal generalization" procedure. Each trial consisted of the presentation of two stimuli (two tones or two visual stimuli, in different conditions), separated by a short gap. The participant's task was to judge whether or not the two stimulus durations were the same (making a YES/NO response), and no feedback was given. To generate the stimulus pairs, a "sample" stimulus was chosen randomly from one of three duration
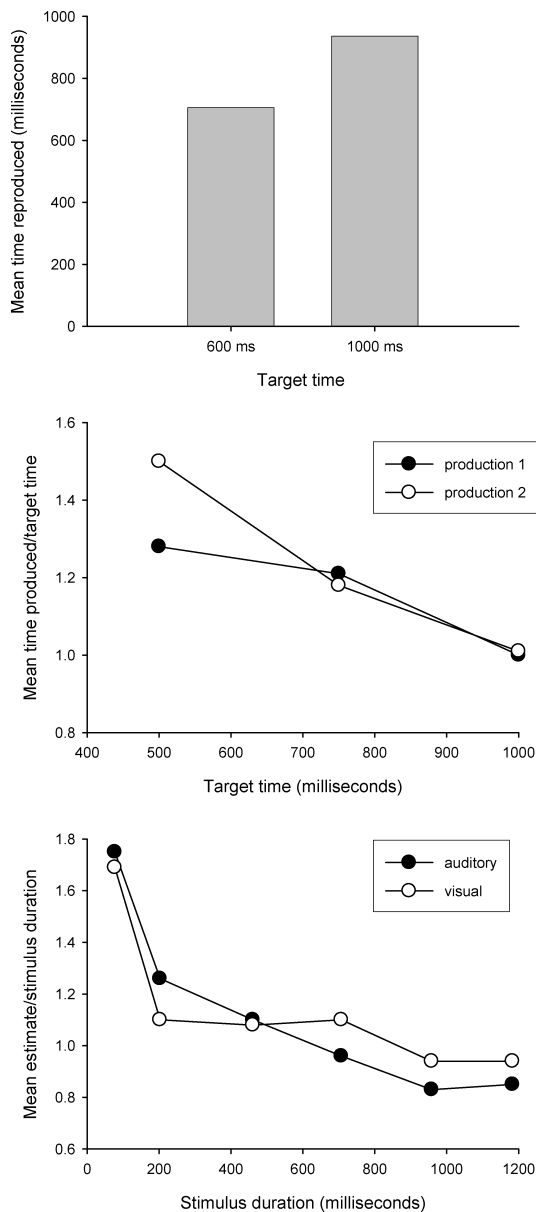
Figure 3. *Illustrations of conformity to Vierordt's law (see text for details). Upper panel: Mean times reproduced when target times were 600 and 1,000 ms. Centre panel: data from two experiments where the target times to be produced were 500, 750, and 1,000 ms. Mean times produced are divided by the target time and plotted against target time value. Lower panel: Data from verbal estimation of the duration of auditory (tones) and visual (squares on a computer screen) stimuli ranging from 77 to 1,183 ms in duration. Mean estimates are divided by real stimulus duration.*

ranges (SHORT: 300–500 ms; MEDIUM: 450–750 ms; LONG: 600–1,000 ms), and a "comparison" duration was generated by multiplying the sample by 0.25, 0.5, 0.75, 1, 1.25, 1.5, and 1.75, with all values being equally likely. Stimulus durations were never repeated, except by chance, but generalization gradients could be drawn in terms of the proportion of YES responses plotted against comparison/sample ratio. Wearden and Bray found good superimposition from all but those conditions where some of the durations were very short (i.e., <100 ms, see their Figures 2 and 4, pp. 295 and 305), and gradient peaks were usually found when the durations were actually equal (comparison/sample ratio = 1.0), indicating mean accuracy.

In contrast to data from "normal" and "episodic" temporal generalization, those collected using the *changing-standard* method of Jones and Wearden (2003, 2004), violate superimposition consistently. Here, a standard is presented at the start of a block (in Jones & Wearden, 2003, 1, 3, or 5 times), then comparison stimuli that are proportions of the standard (i.e., from 25% to 175% of it) are presented, and the participant is required to judge whether each comparison duration is or is not the standard. Then, a new standard is presented, followed by another block of comparisons, and so on. The participants were informed that the standard changed for every comparison block, and standards were picked from a SHORT (300–500 ms) or LONG (600–1,000 ms) duration range. In the three experiments using this method in Jones and Wearden (2003) and two using it in Jones and Wearden (2004) superimposition was hardly ever found, and although peaks of the temporal generalization gradients were sometimes found at the standard, they were often displaced from it. It is unclear why data collected with this method should violate superimposition so consistently, but one speculative suggestion is that the different standards interfere with one another in some way.

Overall, therefore, temporal generalization methods (with the exception of the changing-standard method) provide very strong evidence for both the scalar properties of timing.

Exceptions occur when very short durations are employed and when the tasks compared differ in apparent difficulty of discrimination. Apart from these cases, mean accuracy and (when examined) superimposition have been found in student-age participants, children from 3 to 10 years, and elderly people, in virtually all experiments conducted.

Temporal bisection

The study of temporal bisection in humans predates recent interest in scalar timing in humans (e.g., Bovet, 1968) but in 1991 Allan and Gibbon (1991) and Wearden (1991a, 1991b) independently reported data collected with very similar methods, deriving from Church and Deluty's (1977) bisection procedure for rats. In Wearden's (1991a, 1991b) method people initially received examples of "short" and "long" standards (e.g., unfilled auditory intervals with durations of 200 ms, short, S, and 800 ms, long, L), then received a block of comparison durations (S and L, and stimuli of intermediate duration value), and were required to classify each one in terms of its similarity to S and L. S and L were then presented again, and another block of comparisons was given. No feedback could be given as there is no "right answer" as to how intermediate stimuli should be classified. Allan and Gibbon (1991) used a similar method, but S, L, and intermediate durations were intermixed, and the participant's task was to identify S and L, with feedback given after some responses.

Both studies displayed data in terms of a *psychophysical function*: the proportion of "long" responses (i.e., responses appropriate to L, or judgements that a presented stimulus was more similar to L than S) plotted against stimulus duration. Such functions have an ogival shape, going from near-zero "long" responses when S is presented, to near 100% after L. Given that in these experiments the smallest proportion of "long" responses was produced to S and the largest to L, mean accuracy is implied, so the main focus of interest is on the scalar property of variance. This has been tested in two principal ways. One is to calculate the Weber ratio, a measure of temporal

sensitivity that reflects the slope of the psychophysical function. Another method is to superimpose psychophysical functions from different conditions on the same relative scale, and this time the appropriate relative scale is one that expresses all the durations in a particular condition as a proportion of the *bisection point* (the stimulus duration giving rise to 50% "long" responses) for that condition (Allan & Gibbon, 1991).

Figure 2 shows invented (but typical) results illustrating the shape of the psychophysical function. One of the conditions shown (200/800) involves 200- and 800-ms S and L values; in the other the values are 200 and 400 ms.

Wearden (1991b) studied bisection with S and L values of 200 and 800 ms or 100 and 900 ms. Weber ratios were similar although slightly greater in the condition (100/900 ms) with the larger L/S ratio. Allan and Gibbon (1991) used a range of S and L values, with L up to 3 s, but L/S ratios were always small (2 or less). They reported small (<.10) Weber ratios, but did not report ratios from individual conditions separately. However, data plots (their Figures 6 and 9, pp. 49 and 53) showed good superimposition although there was a suggestion in the data that the psychophysical function from the condition with the largest L/S ratio was flatter than the others (e.g., their Figure 9). For similar results, see Wearden and Ferrara (1995) and Wearden (1995) with the latter experiment reporting data from a "categorical timing" procedure that involves a sort of "tri-" or "quadri-section" of stimulus durations. In both cases, superimposition was generally good (e.g., Wearden, 1995, Figure 3, p. 322).

However, an important qualification to the general presence of scalar properties in bisection data comes from Wearden and Ferrara (1996) who studied bisection with a variety of S and L pairs, with L/S ratios ranging from 2:1 (800/400 ms) to 5:1 (500/100 ms). Inspection of psychophysical functions suggests that the smaller L/S ratios produced steeper slopes (e.g., their Figure 2, p. 31). In a study of bisection in humans with multisecond stimulus durations, Wearden et al. (1997b) likewise found that smaller L/S ratios produced steeper slopes, thus violating superimposition.

The steeper slope in a "more difficult" bisection condition (i.e., one with a smaller L/S ratio) is also illustrated in Figure 2. The values were chosen to produce a steeper slope in the 200/400 condition than the 200/800 one, and the lower panel shows how this causes a failure of superimposition, in the direction of more sensitive timing in the bisection condition with the smaller L/S ratio.

The apparent dependence of temporal sensitivity in bisection on L/S ratio, discussed above, might appear at first sight to contradict some other data from Allan and Gibbon (1991), Allan (2002), and Allan and Gerhardt (2001). In Allan and Gibbon (1991) L/S ratio was varied between conditions, but was always small relative to the values used in some other studies (e.g., from 1.33:1 to 2:1). The data were modelled assuming a constant underlying timing sensitivity, but inspection of psychophysical functions suggests that the smaller ratios produced steeper slopes (e.g., their Figures 4, 5, and 6, pp. 46, 48, and 49). In a study using a wider ratio range, Allan (2002) systematically varied L/S ratio in three experiments using ratios of 2:1, 5:1, and 8:1, and 2:1, 4:1, and 5.8:1. Once again, modelling assumed constant underlying timing sensitivity, and L/S ratio was not the fundamental focus of the article, but inspection of psychophysical functions clearly shows steeper slopes from the lower ratios (e.g., Allan's Figure 1, p. 47).

Allan and Gerhardt (2001) reported some variants of the normal bisection procedure, where the standard procedure (which they called a "no referent" procedure) was contrasted with other methods where the S and L values appropriate for the task in hand were presented on each trial, along with a comparison duration. So, for example, in the "roving referent" condition, different S and L values were randomly generated on each trial, and a comparison duration, which was equal to S or L, or between them, was presented. On average, good superimposition was found whatever the method used. Wearden and Bray (2001) also employed a procedure very similar to the "roving referent" method of Allan and Gerhardt, and their Figure 3 (p. 300) shows that

superimposition was excellent across different absolute duration values. Good evidence for superimposition was also found in Wearden's (2002) analogue of the Gibbon and Church (1981) "time-left" procedure, which has some bisection-like features, so taken together these data suggest that superimposition in bisection may not depend critically on the exact procedure used.

Nor does superimposition depend on using student-age participants. Droit-Volet and Wearden (2001) studied the behaviour of children of 3, 5, and 8 years on bisection with S and L pairs of 1 and 4 s, or 2 and 8 s, and found approximate superimposition (their Figure 3, p. 152), although Weber ratios were larger in the younger children. Droit-Volet and Wearden (2002) studied the behaviour of 3-, 5-, and 8-year-olds with S and L values of 200 and 800 ms or 400 and 1,600 ms and again found good superimposition, particularly in the 8-year-olds (their Figure 2, p. 207).

Overall, bisection provides many examples of superimposition with respect to the absolute values of S and L, from both student-age adults and children. An important exception is that superimposition only holds if the L/S ratio is kept constant between comparisons: If the L/S ratio is changed between conditions, superimposition fails, and steeper psychophysical functions, indicating more sensitive timing, are obtained with smaller L/S ratios.

Discrimination methods

Discrimination methods for studying duration perception long predate modern theories of timing like SET. Normally, participants receive two stimulus durations and have to make some judgement about them—for example, which lasts longer. The relative duration is varied so that the smallest difference that can be reliably detected (a "threshold", a kind of Weber-fraction-like measure of temporal sensitivity) can be measured. Modern procedures usually perform this variation using adaptive methods—that is, where the durations presented on one trial depend on the judgement made on the previous one, although

prearranged trial sequences were often used in earlier work.

For example, in much recent work by Rammsayer and colleagues (e.g., Rammsayer, 1999), the participant receives two white-noise bursts—a standard and a comparison duration (in random order)—and must decide which lasted the longer. On the first trial, the duration difference between the two is very marked, and participants make the correct judgement reliably. Each correct judgement reduces the duration of the comparison by some constant amount, and the adjusted durations are presented on the next trial. This adjustment process continues until an error in made, and then on the next trial the duration of the comparison stimulus is increased (usually by 3 times the amount of the decrease after correct responses). Such a procedure converges on a duration threshold; in the case of steps of 1 down and 3 up, this is the difference detected correctly on 75% of trials.

Differences in procedure between experiments make between-experiment comparisons of these threshold values problematical, but within-experiment comparisons would be appropriate tests of the scalar property of variance. For example, an experiment might vary comparison durations around a short standard or a long one. Is the threshold the same percentage of the standard in both cases (indicating scalar timing), or some different percentage, and how do the threshold percentages change with absolute time values?

The work of Rammsayer and colleagues provides a substantial body of evidence about thresholds and usually employs a short (50 ms) and a long (1,000 ms) standard duration. In a typical study (Rammsayer, 1999) different participant groups received different drug treatments, or placebo. The focus of interest for our purposes is on the relative value of the threshold expressed as a percentage of the standard duration (50 or 1,000 ms). Baseline conditions from Rammsayer (1999, his Table 1, p. 278) show that in all groups the relative threshold was higher for the 50-ms than the 1,000-ms standard. Thus the clear implication of this work is higher relative variance at 50 ms than 1,000 ms, a violation of

superimposition (e.g., Rammsayer, Hennig, Haag, & Lange, 2001; Rammsayer & Vogel, 1992).

Fetterman and Killeen (1992) used a more complex adaptive procedure, as well as a wider range of duration values, up to 5,000 ms. Weber fractions remained roughly constant at values above about 100 ms, but all humans showed marked increases at short duration values, with the shortest value used (25 ms) producing the highest Weber fraction in all individual participants (their Figure 3, p. 87).

Increases in Weber fractions at short stimulus durations are in fact a general theme of earlier studies using discrimination methods, some of which have used extremely short durations. For example, Henry (1948) used tones varying from 32 to 480 ms, and Abel (1972) used unfilled intervals from 0.63 to 640 ms (Abel's Table II, p. 522), and both studies found the highest Weber fractions at the shortest durations.

Kristofferson (1980) in a frequently quoted study in which the author was the only experimental participant, used a discrimination method involving sets of four different empty intervals. The shortest two were associated with a "short" response and the longer two with "long", and 13 stimulus sets with average durations ranging from 100 to 1,480 ms were used. The author conducted 260 experimental sessions, each consisting of 300 trials. For the first 5 sessions, data conformed well to Weber's law, with a CV of about 0.05 for different duration ranges. With more extensive practice, on the other hand, an unusual variance pattern emerged where the variance "quantum" increased very slightly with increasing duration value, only to suddenly "jump" to a much higher value (which then itself slowly increased up until another "jump"). These "quantal" values also appeared to roughly double with doubling of stimulus duration (e.g., 13 ms up to 100 ms, 25 ms at 200 ms, 50 ms at 400 ms, and 100 ms at 800 ms).

Taken overall, discrimination methods with humans provide many violations of the scalar property of variance, although most support the requirement of mean accuracy, at least implicitly.

An important caveat, however, is that many of these studies have used short or ultrashort durations (less than 100 ms), so their data come from a different domain of time perception from that usually studied in research with humans related to SET, such as work on temporal generalization and bisection.

Classical timing procedures: Reproduction, production, and verbal estimation

In the method of *temporal reproduction*, the participant is initially given some sample duration, then must reproduce it either by making some motor response with the same duration, or by adjusting a comparison stimulus so that it has the same perceived duration as the sample.

A common feature of many reproduction experiments is a form of deviation from linear timing known as Vierordt's Law (Vierordt, 1868: The mean of reproductions of short durations is usually longer than the sample, but the reproductions of longer durations is equal to, or shorter than, the sample). The upper panel of Figure 3 shows an example of Vierordt's Law in unpublished data from a condition in which people reproduced intervals of either 600 ms or 1,000 ms. Here, the mean time reproduced exceeded the shorter target time, but was lower than the longer one.

Such results are also found in the classical timing literature. For example, Woodrow (1934) reported mean reproductions of sample durations ranging from 300 to 4,000 ms. The reproduction was longer than the sample at 300 ms, but progressively undershot it as the sample lengthened. Kowalski (1943), similarly, reported data on reproduction of durations ranging from 0.48 to 16.2 s: The mean reproduction was 0.63 s for the shortest duration, but 14.4 s for the longest one (see also Fraisse & Oléron, 1950). In some early studies, the range of durations was such that participants could have used counting-based timing for the longer values, in spite of usually receiving instructions not to do this, but a recent study of reproduction (Wearden, 2003) used sample durations ranging from 500 to 1,000 ms, yet Vierordt-type

effects were still found (his Figure 4, p. 34, see also the present Figure 3).

Although most studies of reproduction of time intervals produce data that violate the linear timing requirement of scalar timing by showing Vierordt-like properties, evidence as to whether the scalar property of variance is violated is mixed. Wearden (2003) found that the CV from noncounting-based reproduction decreased with increasing sample duration in one experiment, but not in another very similar one. Constancy of Weber-fraction-like measures was also reported by Woodrow (1930) where the standard deviation of reproductions was expressed as a fraction of the sample value and varied over values of between 11 and 14% when sample values were from 200 to 2,000 ms, but increased at longer durations.

Rakitin et al. (1998) developed a procedure where people observed a colour change that occurred 8, 12, or 21 s into a trial. They were then were instructed to "center a window of responses around the criterion time" (p. 17), and a feedback procedure was used where the feedback depended on where the response period of the participant was centred. When averaged over trials and participants, the response probability versus elapsed-time functions were of an almost exactly Gaussian shape, with a peak located very close to the criterion time (8, 12, or 21 s), and the response probability functions were superimposed almost perfectly, in exact conformity with the scalar properties of timing. Kladopoulous, Brown, Hemmes, and Cabeza de Vaca (1998) used a rather similar procedure to that of Rakitin et al. (1998), but found evidence for a slight decline in coefficients of variation between 2 and 18 s, with the most marked change between 2 and 4 s, possibly as a result of counting at the longer intervals, which was not controlled.

In the method of *production*, participants are required to make some response or response sequence that lasts some particular length of time, specified in conventional time units (e.g., "hold down this button for one second"). Feedback as to response accuracy may or may not be given. Such techniques share characteristics with some schedules of

reinforcement used in animal laboratories, such as differential-reinforcement-of-low-rate (DRL) or differential-reinforcement-of-response-duration (DRRD). In the former, responses are reinforced only if spaced from previous responses by more than some minimum interresponse time (IRT: the time between successive responses). In the latter, responses such as holding down a lever are reinforced only if their duration exceeds some requirement.

Some studies have tested human participants in variants of such schedule techniques. For example, Zeiler et al. (1987) exposed human adults to a discrete-trial DRL schedule (where participants had to space two successive button presses by durations ranging from 0.5 to 32 s), and people were instructed not to count. A reanalysis of averaged data from this study by Wearden (1991a) showed that behaviour conformed very well to the scalar property of variance, showing a nearly constant CV as the intervals timed varied over a 64-fold range.

In a similar study using a more conventional production method, Wearden and McShane (1988) obtained data from adults on a duration production task (range: 0.5 to 1.3 s). Mean accuracy was obtained from averaged data and from the peak location of fitted curves, and the coefficient of variation remained nearly constant over the range of values used. However, if people are required to produce time intervals without feedback, Vierordt-type effects can be observed (see Wearden, 2003). The centre panel of Figure 3 shows one of these (see also Wearden, 2003). In two different experimental conditions, people produced intervals of either 500, 750, or 1,000 ms, and Figure 3 shows that the mean time produced was longer than the target time at 500 and 750 ms, but approximately accurate at 1,000 ms.

In the technique of *verbal estimation* participants assign verbal labels, in conventional time units such as seconds or milliseconds, to the duration of stimuli or events presented to them. Verbal estimation has recently been used in a number of studies related to the idea that humans possess an internal clock of the sort that SET proposes (e.g., Penton-Voak, Edwards,

Percival, & Wearden, 1996; Wearden et al., 1998; Wearden, Norton, Martin, & Montford-Bebb, in press).

Mean estimates obtained from verbal estimation studies often vary linearly with real time (see Penton-Voak et al., 1996, Figure 4, p. 315, and Wearden et al., 1998, Figure 4, p.110, for examples), but estimates are not simply proportional to real time and instead show Vierordt-like effects. The lowest panel of Figure 3 provides examples from studies in which people estimated the duration of auditory stimuli (tones) or visual stimuli (squares on a computer screen), with real durations varying from 77 to 1,183 ms. Obviously, estimates were longer than the real duration of the stimulus when this was short, but were shorter than it when the longest durations were estimated.

In addition to deviation from the mean accuracy property, coefficients of variation from studies of verbal estimation almost always violate the scalar property of variance, usually by showing a decrease in CV with increasing duration (see Wearden, 1999, Figure 2, p. 10, for examples).

Taken as a whole, data from classical timing procedures, and some modern variants of these procedures, provide very mixed evidence for the scalar property of time. Mean times reproduced, produced, or estimated may show Vierordt-like properties rather than proportional timing, and some procedures like verbal estimation appear to systematically violate the scalar property of variance by showing systematically decreasing CVs as the durations estimated increase. On the other hand, some reproduction and production data can provide very clear evidence for the scalar property of variance.

Violations of empirical scalar timing and their possible causes

We now discuss the four principal violations of empirical scalar timing identified in the above review: very short durations, task difficulty effects, classical timing tasks, and unusual variance patterns. In each case we try to provide some explanation of why such violations occur and try

to reconcile them with underlying scalar representations of time (i.e., with theoretical scalar timing), although, as will be seen, such reconciliation is not always possible at present. To accomplish this reconciliation, we employ some of the details of the SET system (Gibbon et al., 1984), which is briefly expounded below.

SET attributes timed behaviour to underlying processes involving internal clock, memory, and decision mechanisms. The internal clock is assumed to be of a pacemaker–accumulator type, where a pacemaker produces "pulses" or ticks, which are gated to an accumulator, which stores them, via a switch. So, for example, when an event to be timed begins, the switch closes thus establishing the connection between the pacemaker and the accumulator, and at the end of the event the switch opens thus cutting the connection. Accumulator contents thus reflect the "raw material" on which a time judgement can be based. Note that in the example above the switch may have a latency to close and open, and that furthermore this latency may be variable from trial to trial, so switch processes can make a contribution to timed behaviour, as is seen later. SET proposes two memory stores: a working memory, which reflects the contents of the accumulator accurately (in some recent versions of SET this is conflated with the accumulator), and a reference memory of temporal "standards", which depend on the task used. A final stage involves a decision process, which usually entails comparison of accumulator contents with one or more samples from the reference memory, and only after the operation of clock, memory, and decision processes does overt behaviour occur.

Very short durations

Almost all studies that have used durations less than 100 ms have reported increased relative variance (i.e., increased CV or Weber-fraction-like measure) at very short durations compared with longer durations. As Wearden and Bray (2001, pp. 307–308) note, such an effect is exactly what would be expected from a pacemaker–switch–accumulator clock mechanism like that proposed by SET. As mentioned earlier, the operation of

the switch in the clock proposed by SET can contribute variance, by variable latencies to close and open, but we might more generally use the idea of "start/stop" variance—that is, variance due to starting and stopping the timing processes rather than that due to the timing process per se.

Most discussions of such variance (e.g., Wearden et al., 1998) assume that start/stop variance is independent of the duration timed, whereas timing variance is related to the duration timed. Using this argument, start/stop variance will make a much larger relative contribution to the variance observed when timing short intervals than when timing long intervals and at ultrashort intervals may be the principal source of variance. So, the increase of CV seen at very short durations may not represent a serious violation of the underlying scalar property of time, as this property refers to the putative timing variance, rather than the start/stop variance. Indeed, the idea of a constant error added to a basic timing process was incorporated into the "generalized Weber law", which preceded the development of SET (e.g., Getty, 1975).

Task "difficulty" effects

Studies with humans have observed violations of the scalar property of variance in situations in which the "difficulty" of the timing task varies between conditions—for example, Ferrara et al. (1997), Wearden and Grindrod (2003), and Zeiler and Hoyert (1989). As mentioned above, other studies that have not reported this effect (e.g., Allan & Gibbon, 1991) have probably found it in data but left it unremarked.

Ferrara et al. (1997) used the temporal generalization model from Wearden (1992) to model their data and found that the threshold for the decision to respond YES was more conservative when the comparison stimulus duration spacing was smaller. This suggests that the task difficulty effect does not directly violate *theoretical* scalar timing (as the means and variances of the internal representations of the durations did not differ between conditions), even though empirical superimposition is violated. Wearden (2004) also showed that smaller underlying decision thresholds in a bisection model can cause the

psychophysical function to steepen (i.e., timing to become apparently more sensitive), even though the mean and variance properties of the underlying stimuli are constant, suggesting that effects of L/S ratio on bisection performance can also be explained by threshold changes.

Changes in task difficulty may also make participants "pay more attention" to the task involved, but modelling by Ferrara et al. (1997) and Wearden (2004) suggests that the mechanism of this change of attention is not any change in the way that stimuli are perceived, but rather in the decisions made about them. This contrasts with the case where attention is manipulated by requiring participants to perform a concurrent temporal or nontemporal task (e.g., Brown, 1997; Macar, Grondin, & Casini, 1994), where there is evidence that the secondary task interferes with the initial perception of the stimuli judged.

Classical timing tasks

As Wearden (2003) points out, and our review above confirms, data from the "classical" timing tasks of interval production, temporal reproduction, and verbal estimation frequently violate the scalar properties of time in two ways. First, they deviate from mean accuracy, usually in the direction of Vierordt's law, and, second, they frequently violate the scalar property of variance, usually by exhibiting coefficients of variation that decrease as intervals timed get longer (see Wearden, 1999, Figure 2, p. 10). There are cases, however, when mean accuracy and the scalar property of variance is observed with classical tasks or modern variants thereof (Rakitin et al., 1998; Wearden & McShane, 1988; Woodrow, 1930), but violations are common, particularly in situations in which no performance-related feedback is given.

Given the range of applications of SET to human timing (Allan, 1998; Wearden, 2003, 2004) it is perhaps surprising that so little attention has been paid in trying to reconcile SET with "classical" tasks and data. However, Wearden (2003) showed how Vierordt-like properties in temporal reproduction may be compatible with an underlying scalar process. The basic idea was that in a task of temporal reproduction

without feedback, the participant has an accurate representation of the standard, s, as this varies between conditions. The reproduction was modelled by assuming that the participant ran their internal clock until the reading was "close enough" to s (at some time ks), then initiated the response, which took some time, d, (which is independent of s), so the total reproduced time was $ks + d$. For example, if k is 0.6 (i.e., the participant initiated the response at 60% of s), and d is 300 ms, then when s is 500 ms the reproduction would be 600 ms, whereas with s at 1,000 ms, the reproduction would be 900 ms. So, the reproduced time would be longer than the target when the duration to be reproduced was short, but shorter than it when the duration to be reproduced was longer, an effect consistent with Vierordt's law. This example shows that Vierordt-like effects need not be incompatible with underlying linear timing processes, nor incompatible with mean accuracy (in the model s is always represented perfectly accurately).

Production with feedback conforms well to both scalar properties (Wearden & McShane, 1988), but production without feedback does not (Wearden, 2003). This suggests that feedback plays an important role in producing conformity of behaviour to the scalar properties, but exactly how feedback is used in interval production experiments, and in studies of timing in general, remains underresearched. Franssen and Vandierendonck (2002) provide a review of the few earlier attempts to investigate effects of feedback on timing (e.g., Montare, 1985) and discuss some possible mechanisms by which feedback might operate. However, even in their study, effects of feedback were not always dramatic, and the exact mechanism by which feedback exerts effects (when it does) remains somewhat obscure. For further discussion, see Droit-Volet and Izaute (2005), and Wearden and Farrar (in press).

Data from verbal estimation experiments routinely violate the scalar property of variance, although mean verbal estimates are often linear with stimulus duration (Penton-Voak et al., 1996; Wearden et al., 1998, in press). When stimulus durations that do not involve counting

are used (e.g., those below about 1.5 s), participants in verbal estimation studies usually "quantize" their estimates, for example using values ending in "00" in 80–90% or more of cases. Wearden (2006) showed how a computer model simulating this quantization could produce coefficients of variation that declined with increasing duration, even though underlying time representations had the scalar properties of variance.

Another issue that cannot be discussed fully here is why Vierordt-like deviations from linear timing occur with all three "classical" tasks: reproduction, production, and verbal estimation (e.g., Figure 3). Obviously, the processes used in the reproduction model suggested by Wearden (2003) cannot possibly be the cause of Vierordt effects in verbal estimation, where there is no timed motor response, yet such effects occur (see Wearden, 2003, and the present Figure 3). We can only note that Vierordt effects need some systematic treatment by modern theory, but this has not yet been provided. Overall, therefore, the task of reconciliation of data from classical timing tasks like verbal estimation, interval production, and reproduction, with ideas consistent with SET, if possible at all, has hardly begun and may present many significant theoretical challenges.

Unusual variance patterns

We have noted above that relative timing variance often increases when very short durations are employed. In addition, some studies suggest "unusual" patterns of sensitivity, and among these are "step-like" changes in timing variability, or the suggestion that certain absolute time values are timed more precisely than others.

An example of the former effect comes from Kristofferson (1980), who reported deviations from Weberian (i.e., scalar) timing after an abnormally long period of duration discrimination testing (260 sessions), where a series of 20 sessions was devoted to each standard duration (13 standards, ranging from 100 to 1,480 ms). Kristofferson suggested that temporal experience is, in fact, based on "quanta" that were embedded in "noise" early in training, where the quanta

were observable only after extended training had reduced variability (see also Stroud, 1955, and Pöppel, 1978, for earlier similar proposals). If each "step" is representative of a quantum of time, these quanta thus were relative and not absolute, which might not contradict the principles of scalar timing. Obviously, the extensive training needed to demonstrate such quantal effects mitigates against extensive replication, and the reliability of Kristofferson's effects remains to be established, although Killeen and Taylor (2002), in recent theoretical work, provide some bases for the observation of such temporal quanta.

Unusual variance patterns, like those reported by Kristofferson (1976, 1980) seem to depend on very extensive practice, with results obtained early in the experiments usually conforming approximately to Weberian (i.e., scalar) principles. Why practice should have such effects remains mysterious, and a persistent problem is the status of results coming from participants who have had tens of thousands of trials on the same or similar tasks. One view might be that results from such participants are "better" than those obtained from less well-practised participants, with "noise" being reduced so that the real properties of timing are revealed. In contrast, another possibility is that highly practised participants may become sensitive to cues in the experimental procedure (intervals between trials, or other experimental events), which they can use to perform the time judgements, thus rendering their data atypical and perhaps not as revealing of underlying timing processes as those obtained from more normal experimental situations.

Other bodies of results suggest that certain absolute durations are more precisely timed than others, and some data from experiments with humans have recently been reviewed in Grondin (2001). In the early 1990s, Collyer, Broadbent, and Church (1992, 1994) appeared to provide good evidence for the existence of sensitivity "windows" in repetitive tapping in humans. However, a subsequent attempt at replication of these findings in the same laboratory was unsuccessful (Church, Crystal, & Collyer, 1996), and

an "error in the original data collection program" (Church et al., 1996, p. 308) was noted, which seems to call into question the validity of the original findings.

Animal and human timing: Some similarities and differences

In this section, we consider some similarities and differences between aspects of timing found in animals (reviewed in Lejeune & Wearden, 2006) and humans. There are hardly any cases where humans have been tested under conditions that are procedurally similar to those of animals (Fetterman & Killeen, 1992, perhaps come the closest), with experiments using animals and humans differing in the time ranges used, the responses measured, and the way that the behaviour recorded is established and maintained. Even when techniques used with humans are inspired by those used originally with animals the general principle of these "analogue" procedures (Wearden, 1991a) is to pose human participants the same questions about timing that appear to be posed to animals, rather than to aim for procedural similarity between experimental methods used, and these analogue experiments perhaps provide the clearest evidence that at least some timing processes are common to animals and humans (Allan, 1998).

Data collected with temporal generalization and bisection methods suggest many similarities between animal and human timing. Church and Gibbon (1982) tested rats on a temporal generalization procedure where lever presses were rewarded only if a previously presented stimulus had some standard duration (e.g., 4 s), but not when it was shorter or longer. Resulting temporal generalization gradients had two properties in common with those of humans: First, they peaked at the standard duration, and, secondly, they showed superimposition. However, one difference was that the gradients from rats were symmetrical (i.e., with a 4-s standard the response probability was the same after a 3.2-s stimulus as after a 4.8-s one), whereas the gradients obtained from humans are almost always asymmetrical (e.g., see Wearden et al., 1997a, for examples), a difference

that may be attributed to different decision processes in animals and humans (Wearden, 2004).

Likewise, in temporal bisection in rats (e.g., Church & Deluty, 1977, see Lejeune & Wearden, 2006, pp. 1888–1890, for review), the animals are initially trained to press one lever after a short standard and another (the "long" lever) after a longer one; then intermediate durations are introduced, and the response to each is observed without reinforcer delivery. Psychophysical functions from rats resemble those obtained from humans in that the proportion of responses on the "long" lever rises from near zero at the shortest stimulus duration to nearly 1.0 at the longest one, with the increasing function having an ogival shape. The psychophysical functions also superimpose when plotted on the same relative scale, and bisection in rats also shows effects of L/S ratio, with psychophysical functions being steeper when the discrimination is more difficult. However, one difference between bisection and humans and animals is that the bisection point obtained from animals is almost always close to the geometric mean of the short and long standards (the square root of their product), whereas in humans the bisection point is found close to the arithmetic mean, or somewhere between the two. Once again, this animal–human difference can be modelled by differences in decision processes (Wearden, 2004) rather than by any fundamental difference in the way in which time values are represented.

Other clear similarities between timing behaviour in animals and humans that seem likely to reflect similar timing processes come from the discrimination study by Fetterman and Killeen (1992), which directly compared the performance of pigeons and people. Weber-fraction measures of discriminative performance remained constant in both species over a wide range of time values, but increased when the durations timed were very short.

A common procedure used with animals has been to impose some time constraint on the characteristics of a response or group of responses: the so-called temporal differentiation schedules, which bear some resemblance to interval

production procedures used with humans, although animal behaviour is very frequently different, often showing marked deviation from mean accuracy when time requirements for reinforcement are more than a few seconds. However, even in some temporal differentiation experiments similarities between timing in animals and humans can sometimes be noted. For example, Wearden and McShane (1988) found mean accuracy and scalar variance when people produced time intervals ranging from 0.5 to 1.3 s, and Jasselette, Lejeune, and Wearden (1990) found the same result when pigeons produced time intervals from 10 to 70 s, with a perching response.

Another similarity between animal and human timing comes from the data suggesting that certain durations are timed more accurately or with less variance than others, although the similarity of data may be only coincidental, as time ranges used are very different. Crystal (2001) gave rats a single large "meal" at various times after signal onset and found that the timing of the occurrence of this meal was relatively more precise (when measured by activity counts) when the critical time was close to 24 hours than at longer and shorter times, suggesting the involvement of circadian processes. There are no equivalent data from humans with these time ranges although, as discussed above, there is some evidence that unusual variance patterns, where some intervals are timed more sensitively than others, can sometimes be found in data from humans, particularly well-practised humans. Overall, the data from both humans and animals show some evidence that certain absolute time values can in some cases be more precisely timed than others, even when these are not close to 24 hours, but the reliability of the effects, and their explanation, remains unclear.

One consistent difference between data from humans and animals is that timing is less variable in humans (e.g., when assessed in terms of coefficient of variation or Weber-fraction-like measures). Coefficients of variation obtained from human performance are frequently around 0.10–0.15, or even lower, but such values are hardly ever attained by animals, who usually exhibit values twice or more as high. One conclusion to be drawn from this difference is that humans possess a timing system that is more sensitive than that possessed by animals, but it is also possible that attentional differences between the species, or the presence of "untimed" responses of various sorts in animals (i.e., emitted behaviours not controlled by the temporal characteristics of the schedule) play some role in producing the timing sensitivity difference between animals and humans.

CONCLUSIONS

Overall, the present review finds much evidence in favour of conformity of human behaviour to the two scalar properties of timing: mean accuracy and the scalar properties of variance. Systematic violations of empirical scalar timing can be observed, but two of these (very short durations and task difficulty effects) seem easy to reconcile theoretically with underlying scalar representations of time. Unusual variance patterns observed in some experiments cannot so easily be reconciled with scalar timing, but most of these come from situations where participants have performed on the tasks for thousands of trials (with more normal scalar variance being observed early in testing).

From the point of view of human time perception per se, perhaps the most interesting violations of scalar properties in behaviour come from classical timing tasks like production, reproduction, and verbal estimation, where violations seem to be the norm. Workers within the SET framework have always been interested in developing quantitative models of the tasks they use, so it is perhaps surprising that their interest has not turned towards attempts to reconcile the processes of SET with data from classical timing tasks. Such an enterprise, although perhaps difficult, seems likely to be fruitful and offers the possibility of a consistent treatment of data from timing experiments with human participants spanning a century and a half, thus providing a link between classical

studies and modern theory that would be virtually unique in psychology.

REFERENCES

Abel, S. M. (1972). Discrimination of temporal gaps. *Journal of the Acoustical Society of America*, *52*, 519–524.

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

Allan, L. G. (2002). The location and interpretation of the bisection point. *Quarterly Journal of Experimental Psychology*, *55B*, 43–60.

Allan, L., & Gerhardt, K. (2001). Temporal bisection with trial referents. *Perception and Psychophysics*, *63*, 524–540.

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

Bovet, P. (1968). Echelles subjectives de durée obtenues par une méthode de bissection [Subjective scales of duration obtained by a bisection method]. *L'Année Psychologique*, *68*, 23–36.

Brown, S. W. (1997). Attentional resources in timing: Interference effects in concurrent temporal and non-temporal working memory tasks. *Perception and Psychophysics*, *59*, 1118–1140.

Church, R. M., Crystal, J. D., & Collyer, C. E. (1996). Correction of errors in scientific research. *Behavior Research Methods, Instruments, and Computers*, *28*, 305–310.

Church, R. M., & Deluty, M. Z. (1977). Bisection of temporal intervals. *Journal of Experimental Psychology: Animal Behavior Processes*, *3*, 216–228.

Church, R. M., & Gibbon, G. (1982). Temporal generalization. *Journal of Experimental Psychology: Animal Behavior Processes*, *8*, 165–186.

Collyer, C. E., Broadbent, H. A., & Church, R. M. (1992). Categorical time production: Evidence for discrete timing in motor control. *Perception and Psychophysics*, *51*, 134–144.

Collyer, C. E., Broadbent, H. A., & Church, R. M. (1994). Preferred rates of repetitive tapping and categorical time production. *Perception and Psychophysics*, *55*, 443–453.

Crystal, J. D. (2001). Circadian time perception. *Journal of Experimental Psychology: Animal Behavior Processes*, *27*, 68–78.

Droit-Volet, S. (2002). Scalar timing in temporal generalization in children with short and long stimulus durations. *Quarterly Journal of Experimental Psychology*, *55A*, 1193–1209.

Droit-Volet, S., Clément, A, & Wearden, J. H. (2001). Temporal generalization in 3- to 8-year-old children. *Journal of Experimental Child Psychology*, *80*, 271–288.

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.

Droit-Volet, S., & Wearden, J. H. (2001). Temporal bisection in children. *Journal of Experimental Child Psychology*, *80*, 142–159.

Droit-Volet, S, & Wearden, J. H. (2002). Speeding up an internal clock in children? Effects of visual flicker on subjective duration. *Quarterly Journal of Experimental Psychology*, *55B*, 193–211.

Ferrara, A., Lejeune, H, & Wearden, J. H. (1997). Changing sensitivity to duration in human scalar timing: An experiment, a review and some possible explanations. *Quarterly Journal of Experimental Psychology*, *50B*, 217–237.

Fetterman, J. G., & Killeen, P. (1992). Time discrimination in *Columba livia* and *Homo sapiens*. *Journal of Experimental Psychology: Animal Behavior Processes*, *18*, 80–94.

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

Fraisse, P., & Oléron, G. (1950). La perception de la durée d'un son d'intensité croissante [The perception of the duration of a sound of increasing intensity]. *L'Année Psychologique*, *50*, 327–343.

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

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

Gibbon, J., & Church, R. M. (1981). Time-left: Linear versus logarithmic subjective time. *Journal of Experimental Psychology: Animal Behavior Processes*, *7*, 87–108.

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. In J. Gibbon & L. Allan (Eds.), *Timing and time perception* (pp. 52–77). *Annals of the New York Academy of Sciences*, 423.

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

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.

Henry, F. M. (1948). Discrimination of the duration of a sound. *Journal of Experimental Psychology*, *38*, 734–743.

Jasselette, P., Lejeune, H., & Wearden, J. H. (1990). The perching response and the laws of animal timing. *Journal of Experimental Psychology: Animal Behavior Processes*, *16*, 150–161.

Jones, L. A., & Wearden, J. H. (2003). More is not necessarily better: Examining the nature of the temporal reference memory component in timing. *Quarterly Journal of Experimental Psychology*, *56B*, 321–343.

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

Killeen, P. R., & Fetterman, J. G. (1988). A behavioral theory of timing. *Psychological Review*, *95*, 411–422.

Killeen, P. R., & Taylor, T. J. (2002). How the propagation of error through stochastic counters affects time discrimination and other psychophysical judgments. *Psychological Review*, *107*, 430–459.

Killeen, P., & Weiss, N. A. (1987). Optimal timing and the Weber function. *Psychological Review*, *94*, 455–468.

Kirkpatrick, K. (2002). Packet theory of conditioning and timing. *Behavioural Processes*, *57*, 89–106.

Kladopoulos, C. N., Brown, B. L., Hemmes, N. S., & Cabeza de Vaca, S. (1998). The start–stop procedure: Estimation of temporal intervals by human subjects. *Perception and Psychophysics*, *60*, 438–450.

Kowalski, W. J. (1943). The effect of delay upon the duplication of short temporal intervals. *Journal of Experimental Psychology*, *33*, 239–246.

Kristofferson, A. B. (1976). Low-variance stimulus–response latencies: Deterministic internal delays. *Perception and Psychophysics*, *20*, 98–100.

Kristofferson, A. B. (1980). A quantal step function in duration discrimination. *Perception and Psychophysics*, *27*, 300–306.

Lejeune, H., & Wearden, J. H. (2006). Scalar properties in animal timing: Conformity and violations. *Quarterly Journal of Experimental Psychology*, *59*, 1875–1908.

Macar, F., Grondin, S, & Casini, J. (1994). Controlled attention sharing influences time estimation. *Memory and Cognition*, *22*, 673–686.

Machado, A. (1997). Learning the temporal dynamics of behavior. *Psychological Review*, *104*, 241–265.

McCormack, T., Brown, G. D. A., Maylor, E. A., Darby, A., & Green, D. (1999). Developmental changes in time estimation: Comparing childhood and old age. *Developmental Psychology*, *35*, 1143–1155.

Montare, A. (1985). Learning effects of knowledge of results upon time estimation. *Perceptual and Motor Skills*, *60*, 871–877.

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.

Pöppel, E. (1978). *Time perception*. In R. Held, H. W. Leibovitz, & H. L. Teuber (Eds.), *Handbook of sensory physiology* (pp. 713–729). Berlin: Springer.

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

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

Rammsayer, T. H., Hennig, J., Haag, A., & Lange, N. (2001). Effects of noradrenergic activity on temporal information processing in humans. *Quarterly Journal of Experimental Psychology*, *54B*, 247–258.

Rammsayer, T. H., & Vogel, W. H. (1992). Pharmacologic properties of the internal clock underlying time perception in humans. *Neuropsychobiology*, *26*, 71–80.

Staddon, J. E. R., & Higa, J. (1999). Time and memory: Towards a pacemaker-free theory of interval timing. *Journal of the Experimental Analysis of Behavior*, *71*, 215–252.

Stroud, J. M. (1955). The fine structure of psychological time. In H. Quastler (Ed.), *Information theory in psychology* (pp. 174–207). Glencoe, IL: Free Press.

Vierordt, K. (1868). *Der Zeitsinn nach Versuchen* [The experimental study of the time sense]. Tübingen, Germany: Laupp.

Wearden, J. H. (1991a). Do humans possess an internal clock with scalar timing properties? *Learning and Motivation*, *22*, 59–83.

Wearden, J. H. (1991b). 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. (1995). Categorical scaling of stimulus duration by humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *21*, 318–330.

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

Wearden, J. H. (2002). Traveling in time: A time-left analogue for humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*, 200–208.

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). Göttingen, Germany: Hogrefe & Huber.

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

Wearden, J. H. (2006, April). *Mission (almost) Impossible: Modelling the verbal estimation of duration*. Paper presented at the Birmingham meeting of the Experimental Psychology Society, Birmingham, UK.

Wearden, J. H, & Bray, S. (2001). Scalar timing without reference memory: Episodic temporal generalization and bisection in humans. *Quarterly Journal of Experimental Psychology*, *54B*, 289–310.

Wearden, J. H., Denovan, L., Fakhri, M., & Haworth, R. (1997a). Scalar timing in temporal generalization in humans with longer stimulus durations. *Journal of Experimental Psychology: Animal Behavior Processes*, *23*, 502–511.

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., & Farrar, R. (in press). Effects of feedback and calibration on the verbal estimation of the duration of tones. *Acta Psychologica*.

Wearden, J. H., & Ferrara, A. (1995). Stimulus spacing effects in temporal bisection by humans. *Quarterly Journal of Experimental Psychology*, *48B*, 289–310.

Wearden, J. H., & Ferrara, A. (1996). Stimulus range effects in temporal bisection by humans. *Quarterly Journal of Experimental Psychology*, *49B*, 24–44.

Wearden, J. H., & Grindrod, R. (2003). Manipulating decision processes in the human scalar timing system. *Behavioural Processes*, *61*, 47–56.

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., Norton, R., Martin, S., & Montford-Bebb, O. (in press). Internal clock processes and the filled duration illusion. *Journal of Experimental Psychology: Human Perception and Performance*.

Wearden, J. H., Rogers, P., & Thomas, R. (1997b). Temporal bisection in humans with longer stimulus durations. *Quarterly Journal of Experimental Psychology*, *50B*, 79–94.

Wearden, J. H., & Towse, J. H. (1994). Temporal generalization in humans: Three further studies. *Behavioural Processes*, *32*, 247–264.

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

Woodrow, H. (1930). The reproduction of temporal intervals. *Journal of Experimental Psychology*, *13*, 473–499.

Woodrow, H. (1934). The temporal indifference interval determined by the method of mean error. *Journal of Experimental Psychology*, *17*, 167–188.

Zeiler, M. D., & Hoyert, M. S. (1989). Temporal reproduction. *Journal of the Experimental Analysis of Behavior*, *52*, 81–95.

Zeiler, M., Scott, G. K., & Hoyert, M. S. (1987). Optimal temporal differentiation. *Journal of the Experimental Analysis of Behavior*, *47*, 191–200.