THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2004, *57B* (1), 55–77

Double standards: Memory loading in temporal reference memory

Luke A. Jones and J. H. Wearden
University of Manchester, Manchester, UK

Three experiments compared human performance on temporal generalization tasks with either one or two different, and distinct, standard durations encoded. In the first two experiments participants received presentations of two different standards at the beginning of each trial block and were instructed to encode either one or both of them. When instructed to encode one standard they then had to judge whether each of a number of comparison stimuli was or was not that standard. When instructed to encode both they were then tested using just one of the standards but the participants were unaware, at the time of encoding, which standard would later be used as a reference. No marked effect of the number of temporal standards encoded was found. In Experiment 3 participants received either one or two temporal standards and had to use both when two were presented. This manipulation produced flatter generalization gradients when two standards were encoded than when just one was, and modelling attributed this difference mainly to an increase in reference memory variability in the double-standard case. This suggests that the variability of representation of durations in temporal reference memory can be systematically increased by increasing temporal reference memory load.

Scalar expectancy theory (SET: Gibbon, 1977) and its associated information-processing framework (Church, 1984, 1989; Gibbon & Church, 1984; Gibbon, Church, & Meck, 1984) owe their persistence and longevity to their success in explaining both animal and human timing behaviour (Allan, 1998). The information-processing framework itself comprises three major components: a clock process consisting of a pacemaker and an accumulator; a memory process consisting of short-term and reference memory stores; and a comparison process where decisions are made that lead to behaviour output. The reference memory component of the SET model provides a temporal reference for behavioural stability by storing "important" times. In animal studies the content of reference memory will typically be the remembered time of reinforcement; in human studies the reference memory is assumed to contain memories of standard durations important for the task in hand.

By manipulating the parameters of these three components of the SET framework excellent goodness of fit to data has usually been achieved. However, as we have outlined in previous work (Jones & Wearden, 2003), this success is in part based on the flexibility of the

---

Correspondence should be addressed to J. H. Wearden, Department of Psychology, University of Manchester, Manchester, M13 9PL, UK. Email: wearden@psy.man.ac.uk

http://www.tandf.co.uk/journals/pp/02724995.html DOI:10.1080/02724990344000088

components. The fact that the characteristics of the different parts of the SET system can be altered to fit the data (albeit in a plausible and logical way) inevitably leads to problems of falsifiability. In this study we address this issue by using the strategy suggested by Wearden (1999), which is to try to better specify the operating characteristics of the different components by manipulating each of them separately whilst keeping the operation of the others constant.

Such a strategy has already been successful in specifying some of the properties of the internal clock mechanism both in animals (Maricq, Roberts, & Church, 1981; Meck, 1983) and in humans (Droit-Volet & Wearden, 2002: Penton-Voak, Edwards, Percival, & Wearden, 1996). The decision process component of SET has also recently been successfully isolated in humans (Wearden & Grindrod, 2003). Less attention has been devoted to the memory components of the model, although some recent studies have explored timing in the probable absence of reference memory (Allan & Gerhardt, 2001; Rodriguez-Giones & Kacelnik, 1995; Wearden & Bray, 2001).

In previous research (Jones & Wearden, 2003) we attempted to isolate reference memory and explored some of its operating characteristics by manipulating the development of temporal representations. The work presented here continues our exploration of temporal reference memory, but this time our focus is on the question of whether the number of different items in reference memory affects timing behaviour in humans.

Little previous experimental work has been conducted to explore possible effects of increasing the memory load in the reference memory component of the SET information-processing model, and virtually none has explored this question with human participants. However, there are a number of studies that, while not tackling this issue directly, nevertheless indicate some avenues of exploration. Work by Brown and colleagues (Brown, 1997, 1998; Brown, Stubbs, & West, 1992; Brown & West, 1990) is relevant here. These studies explored the performance of participants on temporal reproduction, production, and verbal estimation of multiple target durations. A temporal reproduction task is one in which the participant is required to reproduce a presented stimulus duration—for example, he or she is presented with a tone that lasts 400 ms and is then asked to reproduce that duration by making some motor response (e.g., pressing a key on a computer keyboard). In temporal production the participant is simply instructed to produce a stated duration but without first being presented with an example of the target duration. In a verbal estimation task a person is presented with a stimulus (e.g., a tone), and is asked to judge its duration of presentation, using conventional time units such as seconds or milliseconds. Regardless of the type of task used, the results of all multiple timing procedures of Brown and colleagues showed increased variability in timing with increasing numbers of temporal targets, although the memory used in these sorts of task is perhaps more similar to SET's concept of working memory rather than reference memory.

The usual theoretical explanation of these findings has been in terms of attentional resource allocation. The presence of two or more demanding temporal tasks forces participants to divide attentional resources between two tasks, and, as there is a finite attentional capacity, performance deteriorates as more targets are processed because less attention is allocated to each one (Brown & West, 1990).

The studies of Brown and his colleagues showed that attentional resources are necessary in order for durations to be encoded, and if attentional resources are split between more than one duration then performance is impaired. However, there are some problems involved in extrapolating the findings of these studies to the exploration of temporal reference memory. A major methodological difficulty arises from the fact that target stimuli could overlap with one another, whereas in studies of temporal reference memory putative temporal standards are presented in isolation. As Brown and West (1990) themselves point out, because the presentation times of the stimuli were randomized within each trial, shorter stimuli were much more likely to appear when other targets were also present, so for much or all of the duration of short stimuli other competing stimuli were present, whereas this was not necessarily true for the longer stimuli. Obviously, the presence of simultaneous multiple stimuli is likely to cause problems of encoding, as well as remembering, presented durations, so for explorations of temporal reference memory, a simpler method is needed.

Some studies with animals have investigated performance when more than one time value is associated with reinforced responding. A common method here is the use of mixed-fixed-interval (FI) schedules, where reinforcers are available for the first response occurring at one of the two times after the start of the interval, with nothing signalling to the animal which FI value is in force during a particular interval. Leak and Gibbon (1995) refer to this as *simultaneous* timing but, as Whitaker, Lowe, and Wearden (in press) point out, the timing is not simultaneous in the same sense as that of Brown and West (1990), and the more neutral term *multiple* timing might be preferable.

Whitaker et al. (in press) used 2-valued mixed-FI schedules, where the two components were equiprobable and presented in a random order. They examined the performance of rats on several different FI values when these were presented as single FI schedules and when they were presented as components of mixed FI. For example, in their Experiment 1, rats responded on FI 30 s alone or FI 240 s alone, as well as mixed-FI 30 s 240 s. Gaussian curves were fitted to the response rate versus elapsed time in the interval functions (cf. Lejeune & Wearden, 1991), and these curves could be used to determine the coefficient of variation of the timing process, effectively a measure of temporal sensitivity. The rats timed the single FI schedules more precisely than they did the same time values in the mixed FI, suggesting some sort of "interference" between the two time intervals in reference memory, as is discussed in more detail later.

Olton, Wenk, Church, and Meck (1988) also employed mixed-FI schedules. They found that normal rats were able to time two different stimuli, whether these were presented together or alone. The rats were trained on separate FI 10 s and 20 s schedules, with the duration in force on a particular trial signalled by a tone or light. In the one condition (termed "simultaneous" by Olton et al., 1988), the onset of the 20-s signal was followed by the onset of the 10-s signal after a random delay of 0 to 10 s. On probe trials the rats showed separate response peaks for both durations, identical to those obtained when the intervals were timed separately. However, the focus of this study was the comparison of control and lesioned rats, and it is not clear from the paper exactly what comparisons of performance in single and simultaneous (what we term multiple) conditions were conducted.

There are few studies of humans using the SET framework that have investigated performance when more than one interval is timed, and in addition there are also problems drawing theoretical conclusions from these as they were all conducted without the properties of reference memory being a specific focus of investigation. For example, in their Experiment 1, Malapani et al. (1998) required participants (Parkinson's disease patients and controls) to reproduce two different durations (8 and 21 s) in segregated blocks. Presentations of standards were separated, and nonsimultaneous, so its seems unlikely that initial encoding of the standards was problematic. In their Experiment 2, only Parkinson's patients were used, and here just one duration was encoded and reproduced. There was some indication that encoding two standards increased the variability of the temporal reproductions (but only in Parkinson's patients withdrawn from their normal medication). It is therefore not clear what the effect of encoding two standards would have been on the performance in neurologically intact people, although the suggestion from the Parkinson's patients performing the task when normally medicated is that neurologically intact persons would be able to encode two different durations without much impairment, relative to the encoding of just one.

Another study that used nonsimultaneously presented multiple standards is that of Rakitin et al. (1998). The study used an analogue of the peak procedure (Roberts, 1981) for humans. Comparing timing of one interval (12 s) with timing of more than one (8, 12, and 21 s), showed little difference, but this may be due to the fact that the participants were only given one standard duration per experimental session, so only needed to store a single duration in reference memory at any one time.

One study that may come closer to manipulating reference memory load directly is that of Wearden (1995), which used a categorical scaling technique. With this method, participants were not explicitly presented with standard durations, but had to construct them. On each block the participant received 18 different tone durations in a random order and had to categorize them as belonging to the shortest third, middle third, or longest third of the duration range. In another experiment, the same procedure was repeated, except that this time people were given 24 durations and were required to assign them to four categories. A computer model assumed that the participant constructed the mean duration of the stimuli in each of the three or four categories and then used similarity to these means as the basis for classifying individual durations. In this case, there would be either three or four "items" in reference memory, so variability of performance in these two different conditions could be compared. In fact, contrary to the results in Whitaker et al. (in press), there was no difference in variability between the conditions.

It is obvious from the above that previous work, while sometimes suggestive, does not specifically focus on what seems to us to be the fundamental question of what behavioural changes occur when humans store either a single time value or more than one value in reference memory. In the experiments below, we present distinct, and thus presumably easily discriminable, standards separately and without overlap, thus avoiding any confusion during the phase of encoding into reference memory. In the first two experiments, we use a modified temporal generalization method to investigate the effects of encoding one or two standards, although only one of the standards is used in each block of trials. In the third experiment one or two standards are both encoded and used.

Given the fact that, in all our experiments, standard durations were presented separately there should be no increase in attentional processing at the encoding stage. Therefore, any change in performance under multiple timing conditions will require an alternative explanation. There are a number of possibilities. One is that encoding more than one standard in temporal reference memory produces some kind of interference between the different standards, whereas this interference is absent when only a single standard is used. However, another explanation is that encoding standards in temporal reference memory requires some resource-limited rehearsal, so a second standard interferes with the rehearsal process for the first one and thus disrupts its encoding.

In Experiment 1, we presented two temporal standards, one in the form of a high-pitched tone and the other in the form of a low-pitched tone. Before the presentation of the standards, the participants were instructed that they would be required to compare the duration of subsequently presented tones with *either* the low-pitched tone or the high-pitched tone; thus only one standard needed to be encoded. In another condition they were not instructed as to *which* comparison duration type would subsequently be presented; thus people presumably encoded *both* standards, even though they were subsequently only required to use one of them. The comparison of results from the different conditions may show how increasing memory load affects timing behaviour.

EXPERIMENT 1

To explore the effect of the number of standards encoded on temporal generalization performance, a procedural variant of the normal temporal generalization method was used. In a typical temporal generalization task for humans (e.g., Wearden, 1992) a standard duration is presented a few times (three to five) at the start of the experimental session, and this standard remains in force for the entire session. The participant then receives trials in which comparison durations (a mix of durations shorter than, longer than, or equal in duration to the standard) are presented and must judge whether each duration was the standard or not, with accurate performance-related feedback being given. However, with this method, the temporal representation that the participant is using to make temporal judgements may not be precisely controlled. The fact that the standard remains constant throughout the session may encourage the participant to "construct" the standard using judgements of some of the nonstandard durations (e.g., "the standard is shorter than this one, but longer than this one").

For Experiment 1, we used the "changing standard" temporal generalization method developed by Jones and Wearden (2003). Here, a particular standard is in force only for an experimental block containing a small number of comparison stimuli (but more than one, cf. "episodic" temporal generalization: Wearden & Bray, 2001). The participant was informed of this fact so, presumably, only responded on the basis of the standard/s encoded and currently in force for the block. Using this method, we manipulated the number of the presented standards encoded at the beginning of each block, so the participant encoded either one or both of them and then responded to a small number of comparison durations. If performance is affected by reference memory load then there might be some observable effect of number of standards encoded. We presented two standards on each block even though in half the blocks participants only had to encode one of them. This was because we wanted to keep the stimuli presented constant (except for the varying durations of the standard) so that we could clearly see the effect (if any) of the instructional manipulation of the number of standards encoded. The experiment used standards drawn from two duration ranges (300–500 ms and 600–1000 ms), in order to test the data for the property of superimposition.

Superimposition is the requirement that measures of timed behaviour superimpose when plotted on the same relative scale. Superimposition is a form of conformity to Weber's law and thus implies a timing process with constant underlying sensitivity as the duration timed varies—a requirement of SET. Our method arranges that the comparison durations are multiples of the standard in force for the block, so the quality of superimposition from the different duration ranges can just be examined by plotting the proportion of identifications of a comparison duration as the standard against comparison/standard ratio.

Method

Participants and apparatus

A total of 20 Manchester University undergraduates participated. A Hyundai 386 (IBM-compatible) computer controlled all experimental events. The computer speaker produced the stimuli to be judged, and the keyboard measured participants' responses. The program used to run the experiment and record data was written in the MEL language (Micro-Experimental Laboratory: Psychology Software Tools, Inc), thus providing millisecond accuracy for timing of stimuli and responses.

Procedure

The participants received all trials in one session lasting approximately 15 min. Each block of trials began with the presentation (on the computer screen) of one of three possible instructions, which told the participants that they would (1) be tested on the duration of the high tone (all high tones were 2000 Hz), (2) tested on the duration of the low tone (all low tones were 500 Hz), or (3) tested on the duration of either tone, without specification of which. The participants were then given three presentations of both the high- and low-pitch standard duration tones. Each presentation of one standard was alternated with a presentation of the other standard tone. Between each presentation of a standard there was an interval drawn from a uniform distribution running from 1000–1500 ms, offset to onset.

Each block of trials involved one of eight possible conditions. Participants were instructed that either a high-pitched tone or the low-pitched tone would be the standard, and these were associated with either the short duration range or the long duration range. This constituted four conditions: short/high-pitched; short/low-pitched; long/high-pitched; and long/low-pitched. These cases were the single conditions, since participants only had to encode a single duration (the one corresponding to the pitch instruction). The other four conditions (the double conditions) were identical, except that the participants were instructed that either of the durations might serve as the standard for subsequent comparisons. In these double conditions, the pitch of the comparisons determined which standard was relevant: If the comparisons were low-pitched then the low-pitched standard was to be used; if the comparisons were high-pitched then the high-pitched standard was to be used.

The participants were tested on two blocks corresponding to each of the eight conditions, 16 blocks in all. The order of the blocks was randomized for each participant. The short standard was drawn from a uniform distribution running from 300 to 500 ms; the long standard was always double the length of the short standard, and so ran from 600 to 1000ms. Each block consisted of three presentations of both standards followed by nine test trials.

Following standard presentations, the participant received comparison tones whose duration was the standard duration (whichever one was being tested on that block) multiplied by 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, or 1.6, with the order randomized; each multiplier was placed in an array, and one was picked with equal probability. The standard duration was presented three times as a comparison duration; all other comparison durations were presented once, producing a block of nine comparison stimuli. After each comparison stimulus presentation, the participant judged whether or not the stimulus had the same duration as that of the standard, making a "Y" (Yes) or "No" (No) response on the keyboard. No feedback as to response correctness was given. Each comparison stimulus presentation followed a response to a "press spacebar for next trial" prompt by 750–1250 ms. Following presentation of all nine comparison stimuli, new standard values were randomly generated for the next block and so on. Participants had been informed by previous instructions that the standards would change for each block.

The instructions given to participants are shown in the Appendix.

Results

For the purpose of data analysis the results from these eight different conditions described above were collapsed down into four; single short encoded and tested (single short); single long encoded and tested (single long); two standards encoded, short standard tested (double short); and two standards encoded, long standard tested (double long).

Figure 1 shows temporal generalization gradients, the mean proportion of yes responses (identification of a presented comparison duration as the standard) plotted against comparison/standard ratio. The upper, middle, and bottom panels show data from, respectively, both duration ranges, the short duration range, and the long duration range. In each panel data are shown separately for the single and double conditions.

Consider first analyses of the whole temporal generalization gradients. Inspection of the data in the upper panel of Figure 1 suggests that there was little difference in the overall responding or shape of the temporal generalization gradient between the single and double conditions for either the short or the long standard duration ranges, although the short and long duration ranges produced slightly different proportions of yes responses, particularly to comparison durations longer than the standard. These suggestions were supported by statistical analysis. A repeated measures analysis of variance (ANOVA) used duration range (short or long), and comparison/standard ratio (effectively the duration of the comparison), and number of standards encoded (1 or 2) as within-subject factors. There was no effect of number of standards encoded, $F(1, 19) = 1.87, p = .19$, but there were significant effects of comparison/standard ratio $F(6, 114) = 70.72, p < .001$, indicating that the participants were sensitive to comparison duration, and of duration range, $F(1, 19) = 6.11, p < .05$, indicating that the data from the long and short duration ranges produced different proportion of yes responses overall. There was no significant interaction between number of standards encoded and comparison/standard ratio, $F(6, 114) = 0.68, p = .67$, nor between duration range and number of standards, $F(1, 19) = 0.97, p = .34$, but there was a significant interaction between duration range and comparison/standard ratio, $F(6, 114) = 3.40, p < .01$, indicating a failure of superimposition of temporal generalization gradients from the short and long duration ranges. The three-way interaction between number of standards, duration range, and comparison duration was not significant, $F(6, 114) = 1.08, p = .38$.

Discussion

Experiment 1 failed to find any effect of the number of standards encoded on the overall level of yes responses, or the shape of the temporal generalization gradient. There may be several reasons why an effect of number of standards was not obtained. It may be that encoding two standards rather than one is not cognitively demanding enough and does not use significantly more processing than encoding one. Alternatively, it may be that increasing memory load by encoding two standards *is* significantly more cognitively demanding than encoding one but that our temporal generalization task was not sufficiently sensitive to reveal this. We therefore decided to repeat Experiment 1 with a number of changes.

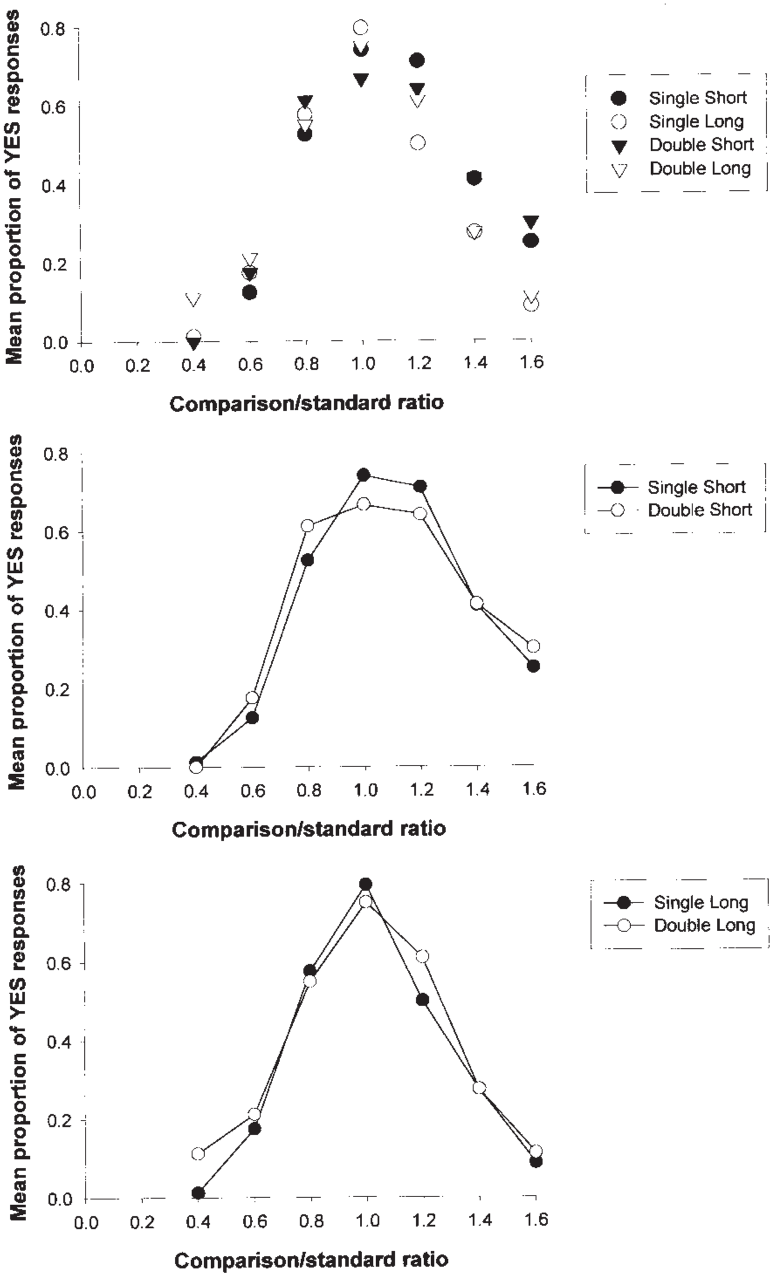

Figure 1. Temporal generalization gradients (proportion of identifications of a duration as the standard, i.e., yes responses, plotted against comparison stimulus durations expressed as fraction of the standard duration) for both conditions (single and double) of Experiment 1. Upper panel: data from both duration ranges (short and long). Centre panel: data from short duration range. Bottom panel: data from long duration range.

For Experiment 2, we increased the difficulty of the task, by reducing the spacing of the comparison durations around the standard, and in Experiment 2 the comparison/standard ratios were 0.5, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, or 1.5. The two extreme values were included to provide participants with relatively "easy" discrimination and thus maintain their motivation to perform the task.

EXPERIMENT 2

Method

Participants and apparatus

A total of 16 Manchester University undergraduates participated, and the apparatus was that used in Experiment 1.

Procedure

Each participant received a single experimental session lasting approximately 20 min. The procedure was identical to that of Experiment 1 except for the spacing of the comparison durations around the standard in the test trials of each block. Following standard presentation, the participant received comparison tones whose duration was the standard duration (whichever one was being tested on that block) multiplied by 0.5, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, or 1.5, with the order randomized. The standard duration was presented three times as a comparison duration, all other comparison durations were presented once, and thus each block consisted of 11 test trials. Each block of trials was one of eight possible conditions described in Experiment 1 and twice on each, making 16 blocks in all. The order of the blocks was randomized for each participant.

Results

As in Experiment 1, for the purpose of data analysis the results from the eight conditions were collapsed down into four conditions: single short standard (single short); single long standard (single long); two standards encoded, short standard tested (double short); and two standards encoded, long standard tested (double long).

Figure 2 shows temporal generalization gradients: the mean proportion of yes responses (identification of a presented comparison duration as the standard) plotted against comparison/standard ratio. The upper, middle, and bottom panels show data from both duration ranges, the short duration range, and the long duration range respectively. In each panel data are shown separately for single and double conditions.

Consider first the analyses of the whole temporal generalization gradients. A repeated measures ANOVA used duration range (short or long), and comparison/standard ratio (the duration of the comparison), and number of standards encoded (1 or 2) as within-subject factors. There was an effect of number of standards encoded, $F(1, 15) = 7.15$, $p < .05$, indicating that the number of standards did affect overall level of responding. There was also a significant effect of comparison/standard ratio, $F(8, 120) = 34.83$, $p < .001$, indicating that the participants were sensitive to comparison duration. There was no interaction between number of standards encoded and comparison/standard ratio, $F(8, 120) = 1.35$, $p = .26$, suggesting that the number of standards encoded did not significantly affect the shape of the temporal

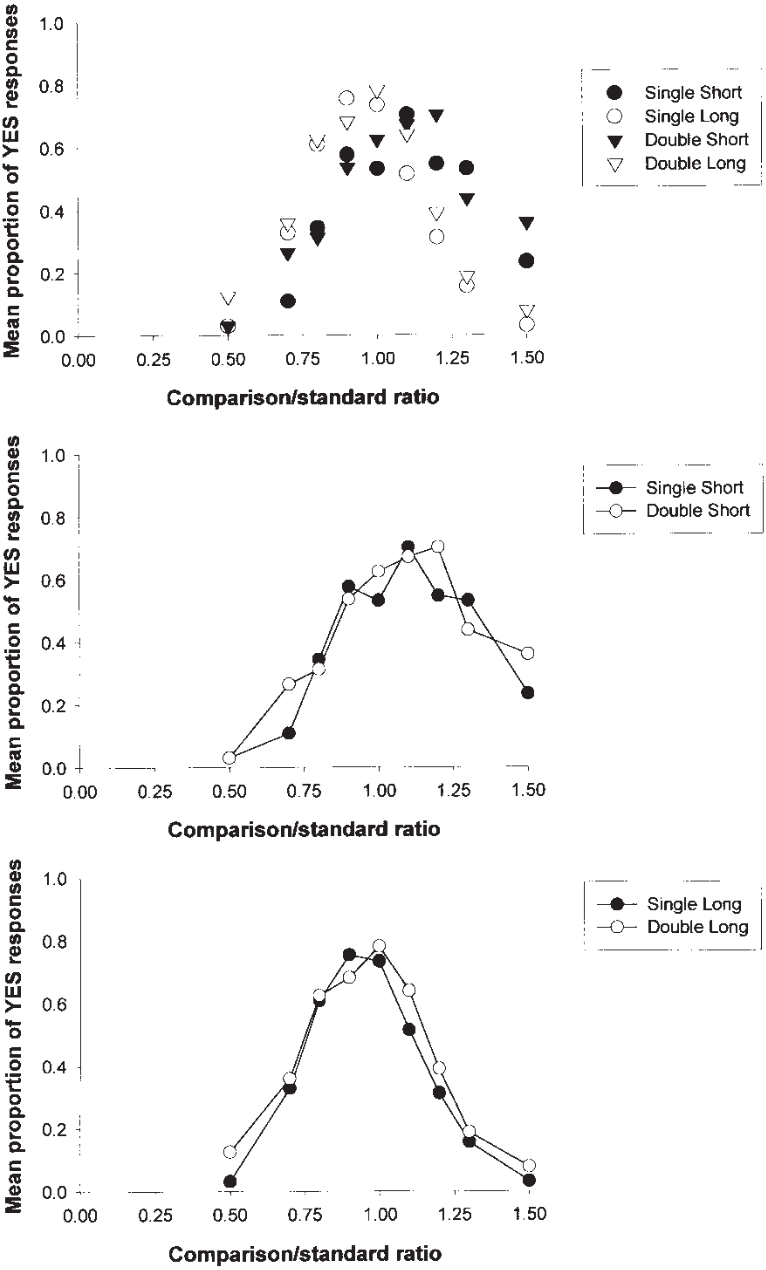

Figure 2. Temporal generalization gradients for both conditions (single and double) of Experiment 2. Upper panel: data from both duration ranges (short and long). Centre panel: data from short duration range: Bottom panel: data from long duration range.

generalization gradients. There was also no significant main effect of duration range, $F(1, 15) = 0.62$, $p = .44$, indicating that the duration range of the standard encoded did not affect the overall level of responding, but there was a significant interaction between duration range and comparison/standard ratio, $F(8, 120) = 11.72$, $p < .001$, indicating that the temporal generalization gradients from the short and long duration ranges differed significantly in shape and thus did not superimpose. The three-way interaction between number of standards, duration range, and comparison duration was not significant, $F(8, 120) = 0.76$, $p = .58$.

Data from the short and long duration ranges were then analysed separately. Consider the short duration range first. A repeated measures ANOVA used number of standards encoded (1 or 2) and comparison/standard ratio as within-subject factors. There was no effect of number of standards encoded, although the effect did approach significance, $F(1, 15) = 3.50$, $p = .08$, but there was a significant effect of comparison/standard ratio, $F(8, 120) = 15.98$, $p < .001$. The interaction between number of standards and comparison/stimuli ratio was not significant, $F(8, 120) = 1.50$, $p = .205$.

The same analysis was conducted on data from the long duration range. The effect of number of standards encoded approached significance, but did not reach it, $F(1, 15) = 4.04$, $p = .06$, but there was a significant effect of comparison/standard ratio, $F(8, 120) = 38.42$, $p < .001$. The interaction between number of standards encoded and comparison/standard ratio was not significant, $F(8, 120) = 0.52$, $p = .738$.

Discussion

Experiment 2 replicated the principal findings of Experiment 1 in that there was no significant effect of encoding one or two standards on most features of the temporal generalization gradient. However, there was a significantly higher overall proportion of yes responses when two standards are encoded compared to one, but when the short and long duration ranges were analysed separately this effect only approached significance. As in Experiment 1, data from the short and long duration ranges did not superimpose: Although failures of superimposition are quite rate, the changing standard generalization method produced this effect in the only previous study in which it has been used, that of Jones and Wearden (2003).

EXPERIMENT 3

The previous two experiments failed to yield any marked difference in temporal generalization gradients after the encoding of one versus two standards. There are a number of possible explanations for this. The first involves interference. It may be the case that in the single conditions, where the participants are told to attend to only one of the presented standards, the second irrelevant standard interferes with rehearsal and encoding of the first. Thus there may be no real difference between the single and double conditions due to the second standard that has to be encoded on the double blocks and ignored on the single blocks, impairing performance on both tasks.

A second possible explanation is that single and double conditions are influencing behaviour differently but that the duration discrimination task was not difficult enough to reveal any significant differences in performance. Some slight support for this idea comes from Experiment 2, as when the difficulty of the task was substantially increased compared with Experiment 1, there was a just-significant overall difference in performance between single and double conditions, although this effect was small and not significant when the short and long duration ranges were analysed separately.

The last and perhaps the most interesting possible reason for a lack of significant effects is that the results are telling us something important about reference memory. The lack of effect obtained in Experiments 1 and 2 may be rooted in how reference memory is defined. If we examine again more closely what we suppose the participants are doing when given the double condition in the previous experiments, the importance of this issue becomes clear. In the double conditions the participant does not know which of the two presented standards they will subsequently be tested on. Therefore we expected that both standards would be encoded into reference memory, even though only one of the standards would be used for all the subsequent trials of that particular block. Alternatively, it could be the case that both presented standards enter working memory, but only the "useful" standard is subsequently transferred to reference memory after the first comparison trial, when the participant becomes aware of which standard is to be tested for that particular block. So, in fact, both our single and double conditions involve a single item being stored in reference memory.

This raises the important issue of exactly what is meant by *reference memory*. There is some confusion and disagreement in the literature on this issue, although the conflicts between alternative definitions are not explicit, as in previous studies reference memory has usually not been the particular focus of interest. There appear to be two definitions of reference memory, and with respect to our experiments these have quite different implications. Reference memory can be defined either in terms of *reinforcement and critical durations*, or in terms of *utilization*.

The method of investigation that we have used thus far is built upon a definition of reference memory in terms of reinforcement and critical durations, because our experiments are based on human analogues of animal experiments. We assume that because the standards are identified as such and because either of them could be important for the subsequent trials of a particular block, then the participant will necessarily encode them into reference memory, a position consistent with some previous definitions, for example:

> The reference memory contains representations of "important times", such as those associated with reinforcement in animal experiments, or identified in some way as standard in experiments with humans. (Wearden & Bray, 2001, p. 291)

and

> Reference memory is assumed to be composed of a distribution of values based on counts that had been in the accumulator. . . . When reinforcement is received, it is assumed that the current count in the accumulator (an integer) is multiplied by a real number (k*), and that this number is put into the distribution of values in reference memory. (Church, 1997, p. 51)

Thus, if we assume that, in experiments with humans, identification of a duration as a standard is analogous to associating a particular duration with reinforcement in animal studies (and this operation is sufficient to guarantee that the duration is stored in reference memory), then we might have expected to find a significant difference between the single and double conditions. This expectation is based on the assumption that the identification of a duration as a standard would be sufficient for it to be encoded and thus increase memory load, regardless of whether both standards are actually used to make temporal judgements.

The other, somewhat conflicting, definition of reference memory is concerned with how different types of memory are used, and it may actually predict absence of a significant difference between the single and double conditions, or the very small and rather ambiguous effect, found in our Experiments 1 and 2, thus:

> Working memory contains information relevant to a single trial, whereas reference memory contains information relevant to all trials. (Meck, Church, & Olton, 1984, p. 3)

and also

> retention of recent events of transient importance (working memory. . . . the retention of information that is of long term importance (reference memory). (Meck & Church, 1987, p. 457)

If we apply this single- versus multiple-trial definition to our experiments then it is clear that one of the standards in our Experiments 1 and 2 is not relevant to even a single trial, let alone all trials. Thus participants in the double condition may not be encoding two standards into reference memory at all, and consequently a significant difference between the single and double conditions would not be expected. In temporal generalization, items may be stored in reference memory only if they are going to be subsequently used, for example, as the basis for decisions in the next few trials.

The essential point here is that because the irrelevant standard on the double blocks is never tested we cannot be sure whether it is encoded into reference memory from working memory or not. Is it really conceivable that being told that a particular tone duration is (or may be) a standard is equivalent to the reinforcement necessary for storage in reference memory according to the Church (1997) definition? To hold both standards in working memory and transfer the relevant one to reference memory on the first trial may be a more cognitively efficient strategy. Thus the critical issue may be whether the transfer of accumulator/working memory contents into reference memory is dependent upon simple presentation of potential standards identified as such by instructions, or whether it depends on function; thus to be transferred to reference memory a standard "must be relevant for all trials" (Meck et al., 1984), or at least for some of them

We conducted a third experiment to address some of these issues. In Experiment 3 there were three different types of task: a temporal generalization task with a short duration standard; a temporal generalization task with a long standard duration; and a double generalization with both long and short standards. Experiment 3 used a "normal" temporal generalization procedure (e.g., Wearden, 1992) so the standard durations remained constant throughout the experiment at 400 ms (short duration), and 700 ms (long duration). In the double generalization task the participants were given four presentations of a long standard and four presentations of a short standard, and they were instructed that both would subsequently be used. Participants were then presented with comparison durations, of which half corresponded to the short standard and half to the long standard. As in the previous experiments the comparison and standard durations were identifiable by the tone pitch. Another difference from previous experiments was that feedback as to performance correctness was given after each response.

Experiment 3 thus forced participants to use both two standards in the double condition. Therefore if a significant difference were found between the single and double conditions this would support the definition of reference memory given by Meck et al. (1984) and have implications for the analogies drawn between human and animal experiments.

Method

Participants and apparatus

A total of 24 Manchester University undergraduates and postgraduates participated. The apparatus was that used in Experiments 1 and 2.

Procedure

Participants were randomly assigned to two separate groups (12 to Set A and 12 to Set B). Each participant then carried out each of the three tasks corresponding to their set allocation. There were six possible orders for the three tasks in each group, and two participants from each group were assigned to each of these six possible orders.

The three tasks were single temporal generalization with a 400-ms standard, single temporal generalization with a 700-ms standard, and double generalization with 400- and 700-ms standards. The tasks in Set A and Set B differed only in the pitch of tone used for the stimuli. In order to counterbalance for any difference caused by the pitch of the standards Set A used a high-pitch (2000-Hz) tone for the 400-ms standard and a low-pitch (500-Hz) tone for the 700-ms standard in both the single and the double conditions. In Set B the pitches were reversed: thus high pitch for 700-ms standard and low pitch for 400-ms standard.

Single standard conditions. The procedure was identical for all the single conditions in both Set A and Set B, with only the pitch and duration of the standard (and the subsequent comparison tones) differing.

Each task began with four presentations of the standard duration tone. The presentations were separated by a random gap drawn from a uniform distribution running from 2000 ms to 3000 ms. All subsequent stimuli followed a participant's spacebar press prompted by "Press spacebar for next trial" presented on screen. After each comparison stimulus was presented the participants were asked, "Was that the standard tone? Press Y (Yes) or N (No) keys". After each response accurate feedback was presented on the computer screen (i.e., "Correct. That was the standard tone", "Incorrect. That was not the standard tone", "Correct. That was not the standard tone" or "Incorrect. That was the standard tone") for 200 ms.

The comparison durations were equal to the standard multiplied by 0.25, 0.5, 0.75, 1, 1.25, 1.5, or 1.75, with the order randomized. The standard tone was presented three times in each block. Thus each block consisted of nine trials, and there were five blocks in all.

Double standard condition. The procedure was identical for the double conditions of Set A and Set B, with only the pitch and duration of the standard (and the subsequent comparison tones) differing. The procedure was essentially the same as that for single generalization except that each task began with four presentations of each of the two standard durations. Each presentation of one standard was alternated with a presentation of the other standard tone. The duration of the gap between presentations of the standards was an interval drawn from a uniform distribution running from 2000–3000 ms offset to onset.

The comparison durations were equal to the standard multiplied by 0.25, 0.5, 0.75, 1, 1.25, 1.5, or 1.75, with the order randomized. If the comparison was a high-pitched tone then the participants were required to compare its duration to that of the high-pitched standard, and conversely if low-pitched they compared the comparison duration to low-pitched standard. Each standard tone was presented three times in each block. Thus each block consisted of 18 trials, and five blocks were given in all.

Results

For data analysis the conditions were collapsed across pitch; thus four conditions were of interest: performance on 400-ms standard in the single condition (single short); 700-ms standard in the single condition (single long); 400-ms standard in double condition (double short); and 700-ms standard in the double condition (double long).

Due to a computer error, data from the single standard conditions with high-pitched tone were lost for one participant, so the data of the 23 remaining participants were analysed. Figure 3 shows temporal generalization gradients: the mean proportion of yes responses (identification of a presented comparison duration as the standard) plotted against comparison/standard ratio. The upper, middle, and bottom panels show data from both standard durations, short standard duration, and long standard duration, respectively. In each panel data are shown separately for the single and double conditions.

Consider first the analyses of the whole temporal generalization gradients. Inspection of the data in Figure 3 suggests a difference in the overall responding and shape of the temporal generalization function between the single and double conditions for both the 400-ms and the 700-ms standards. This suggestion was supported by statistical analysis. A repeated measures ANOVA used standard duration (long or short: 400 ms or 700 ms), and comparison/standard ratio (effectively the duration of the comparison), and number of standards encoded (1 or 2) as within-subject factors. There was an effect of number of standards encoded $F(1, 22) = 15.66$, $p < .01$, indicating that the number of standards encoded affected the overall level of responding. There was also a significant effect of comparison/standard ratio, $F(6, 132) = 221.35$, $p < .001$, showing that the participants were sensitive to comparison duration. There was also a significant interaction between the number of standards and comparison/standard ratio, $F(6, 132) = 4.56$, $p < .01$, showing that the number of standards encoded significantly affected the shape of the temporal generalization gradients. There was, however, no significant main effect of standard duration, $F(1, 22) = 1.69$, $p = .21$, indicating that the duration range had no effect on the overall level of responding. There was also no significant interaction between duration range and comparison/standard ratio, $F(6, 132) = 0.62$, $p = .53$, indicating that the temporal generalization gradients from the short and long standard duration did not differ significantly in shape, thus superimposed. The three-way interaction between number of standards, duration range, and comparison duration was not significant, $F(6, 132) = 0.87$, $p = .45$.

Next, data from the short and long standard conditions (middle and lower panels of Figure 3) were analysed separately. Taking the short standard first (single short and double short), a repeated measures ANOVA used number of standards encoded (1 or 2) and comparison/standard ratio was within-subject factors. There was a significant effect of the number of standards encoded, $F(1, 22) = 19.16$, $p < .001$, and a significant effect of comparison/standard ratio, $F(6,$

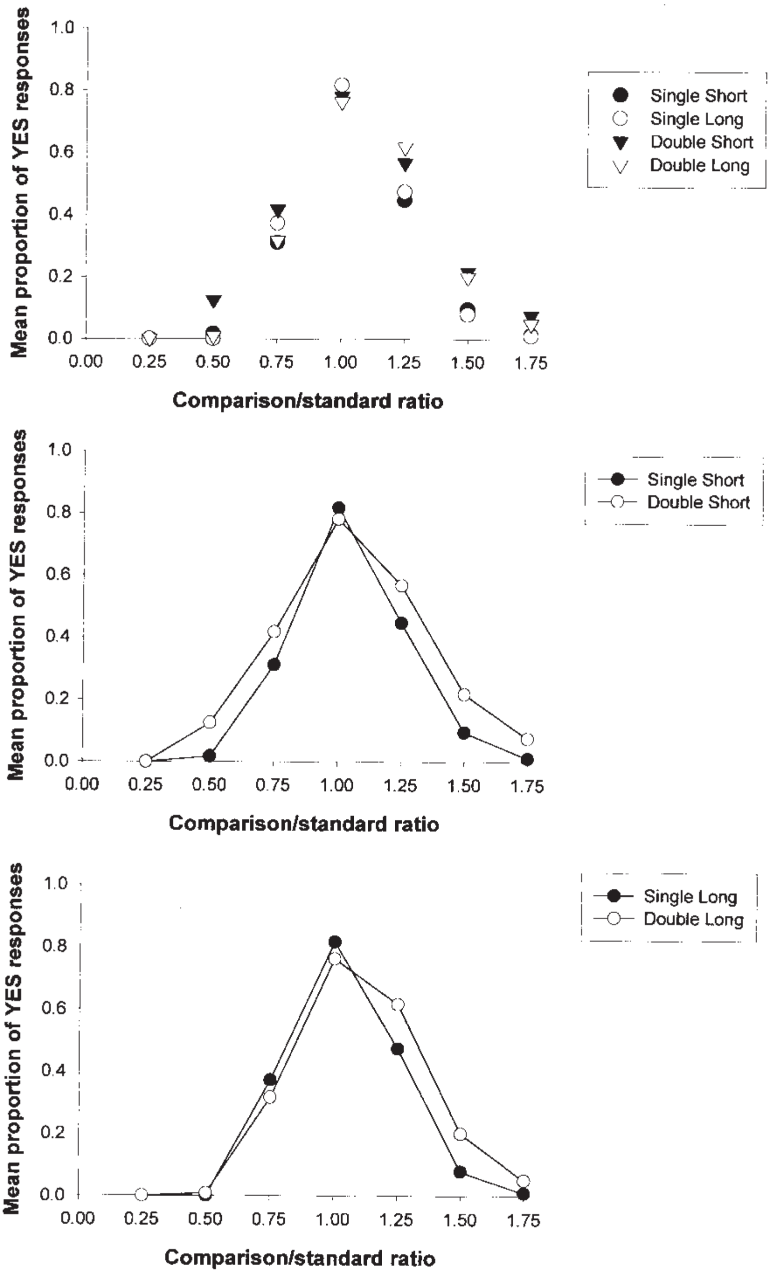

Figure 3. Temporal generalization gradients for both conditions (single and double) of Experiment 3. Upper panel: data from both duration ranges (short and long). Centre panel: data from short duration range. Bottom panel: data from long duration range.

132) = 93.15, $p < .001$. There was no significant interaction between number of standards and comparison/standard ratio, $F(6, 132) = 1.17$, $p = .16$.

The same analysis was conducted for the long standard conditions (single long and double long). There was a significant effect of number of standards encoded, $F(1, 22) = 4.67$, $p < .05$, and a significant effect of comparison/standard ratio, $F(6, 132) = 115.96$, $p < .001$, as well as a significant interaction between number of standards encoded and comparison/standard ratio, $F(6, 132) = 3.48$, $p < .01$.

Discussion

In Experiment 3 one or two standards were encoded, as in Experiments 1 and 2, but this time when two were encoded both were tested. This produced significant effects of the number of standards encoded on both the overall level of responding and the shape of the temporal generalization gradient. In Experiments 1 and 2 the only effect of the number of standards encoded was a small increase in overall level of responding in Experiment 2.

Because there was always a second standard present in the single conditions of Experiments 1 and 2, and not in the single condition of Experiment 3, it could be argued that the reason we obtained an effect between single and double conditions in Experiment 3 and not in Experiments 1 and 2 was because of a rehearsal interference effect. However, this seems unlikely, as when the single conditions of Experiment 1 and 2 (where the first standard presented is the one that participants are told to encode—Conditions 1 and 3 of Experiment 1) and that of Experiment 3 are compared in terms of correct identification of the standard when presented, there was no significant difference between them, Kruskal–Wallis, $\chi^2_{(2)} = 0.58$, $p = .750$. In other words, the presence of a second irrelevant standard at the time of encoding did not cause a rehearsal interference effect in Experiments 1 and 2. Therefore the significant differences between the single and double conditions of Experiment 3 are probably not attributable to rehearsal interference.

Having found an effect of increasing memory load on temporal generalization performance in Experiment 3, we used computer modelling to ascertain the source of the effect. We simulated performance using the model derived by Wearden (1992) for temporal generalization in humans, derived from an earlier model by Church and Gibbon (1982), called the modified Church and Gibbon (MCG) model. This has previously been shown to provide a good fit to temporal generalization gradients obtained from humans by Wearden (1992), Wearden and Towse (1994), and Wearden, Denovan, Fakhri, and Haworth (1997). The MCG model produces a yes response when $|s^* - t|/t < b^*$, where s^* is the sample drawn from reference memory, t is the just-presented duration (comparison), and b^* is a threshold value, which is variable from trial to trial. The value s^* is a sample drawn from a Gaussian distribution with an accurate mean, s, equal to the real-time value of the standard, and some coefficient of variation, c. Larger c values indicate greater reference memory variability. The value b^* is also drawn from a Gaussian distribution with mean b, and for the present simulations, and most others (e.g., Wearden et al., 1997), the standard deviation of this distribution is kept constant at $0.5b$. The value t, the comparison duration, is assumed to be timed without error. Different temporal generalization gradients could be simulated by differences in c, reference memory variability, differences in b, the decision threshold, or both. Droit-Volet, Clément, and

Wearden (2001) show in their Figure 1 (p. 273) how changes in the values of *c* and *b* affect predicted generalization gradients.

The two parameters *c* and *b* were varied over a wide range, and the conditions of Experiment 3 were simulated with 10,000 trials at each comparison duration value. The parameters were varied until a best fit in terms of mean absolute deviation between simulated and experimental data from Experiment 3 was found. Table 1 shows the parameter values obtained, and Figure 4 shows the fit of the model output (unconnected lines) to data points (filled circles) for the four conditions (single short, double short, single long, double long) modelled.

Inspection of the parameter values obtained from the fit of the MCG model suggests several conclusions. First, the small mean absolute deviations obtained (*MAD* in Table 1) showed that the MCG model fitted the data well in all cases. Second, the coefficient of variation of the memory representation of the standard, *c*, was the same for fits of the single short and single long conditions, and very similar for the double short and double long conditions, a theoretical confirmation of the property of superimposition of short and long conditions discussed earlier. Third, to accommodate the behavioural changes between the single and double conditions (single short vs. double short, single long vs. double long), the main parametric change needed was an increase in *c*, reference memory variability, although the decision threshold value, *b*, also increased between single and double conditions. This suggests that the principal effect of encoding two standards in reference memory compared with one was an increase in the "fuzziness" of the representations of the standard in reference memory.

An anonymous reviewer queried whether the behavioural differences between the single and double conditions could be simulated well by varying only one of the parameters of the MCG model (*c* or *b*) while keeping the other constant. We tested this by attempting to simulate data from the double conditions using one of the parameters from the corresponding single condition (short or long), while varying the other. When fitting the double short condition with $c = .13$ (the best fitting value from the single short fit), the smallest *MAD* attainable was .05, and when fitting the double short data with $b = .2$, the smallest *MAD* was .07. Both these

TABLE 1
Best fitting parameter values for fits of the modified Church and Gibbon model[a] to data from Experiment 3

	Parameter		
Condition	*c*	*b*	*MAD*
Single short	.13	.20	.02
Double short	.19	.25	.03
Single long	.13	.21	.03
Double long	.18	.24	.02

Note: *c* is coefficient of variation of reference memory distribution; *b* is decision threshold; *MAD* is mean absolute deviation (sum of absolute differences between predictions of model and data points divided by 7, the number of data points).

[a] Wearden (1992).

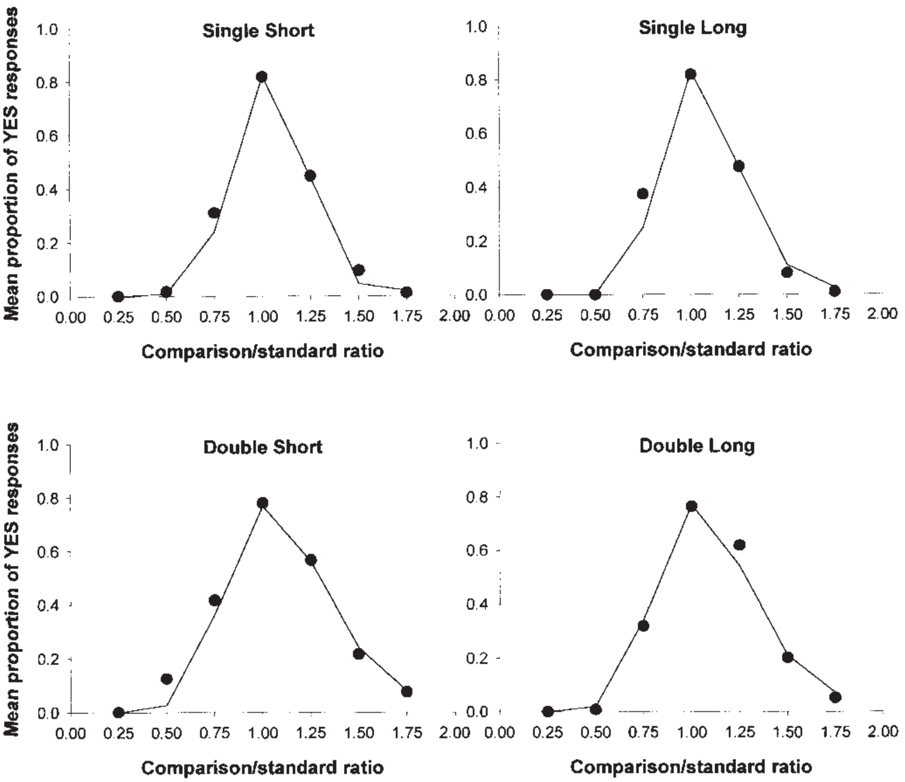

Figure 4. Temporal generalization gradients from Experiment 3 (filled circles) and values predicted by best fitting MCG model (solid line). Upper left panel: single short condition. Upper right panel: single long condition. Lower left panel: double short condition. Lower right panel: double long condition.

values indicate worse fits than those shown in Table 1, when both *c* and *b* could vary. We fitted data from the double long conditions in the same way. Keeping *c* constant (at .13) produced nearly as good a fit (*MAD* = .03, with *b* = .25) as varying both *c* and *b*, but varying *c* alone produced a worse fit (*MAD* = .04). Overall, therefore, varying both *c* and *b* in the MCG model between single and double conditions produced better fits than varying only one of these, although the differences were most marked for the short conditions. Inspection of the data in Figure 3 shows, however, that this was only to be expected, as the differences between single and double conditions were more marked for short durations than for long ones.

GENERAL DISCUSSION

From the results of Experiments 1 and 2 it would appear that encoding two standards, but only being required to use one, does not increase the load on reference memory in any very marked way. Thus, taking the definition of reference memory as being simply the encoding of an "important time" does not seem to account for the fact that just encoding two standards does not lead to any strong effect of memory load. However, when the experimental conditions conform to the definition of reference memory as encoding an item that is useful for all trials (or at least more than one), as in Experiment 3, then we did observe a clear memory loading effect. An essential defining characteristic of a temporal reference memory would thus appear to be the encoding of a standard that is useful for more than one trial, otherwise there would have been no difference between Experiments 1 and 2 and Experiment 3. When this definition is met performance exhibits greater behavioural variability when two standards are encoded and used as opposed to one.

This result parallels others obtained from rats on mixed-FI schedules by Whitaker et al. (in press). In their study, some experiments involved rats receiving exposure to an FI value alone, then having the same FI value as one component of a 2-valued mixed-FI schedule. A complicating factor in Whitaker et al.'s results was that the behaviour of the rats was different if the two standard durations making up the mixed-FI schedule had similar values (ratio less than 3:1) or more different values (ratios greater than 3:1). In the latter case, the animals appeared to have separate representations of the two potential times of reinforcement. As mentioned earlier, response rate versus elapsed time functions produced by the rats were analysed by fitting single Gaussian curves to single FI schedules and the sum of two independent Gaussian curves to data from mixed-FI schedules. The Gaussian curves yielded a coefficient of variation statistic that indicates the acuity of temporal control, and when data from single FI schedules, and the same FI values appearing as a component in a 2-valued mixed-FI schedule, were compared, and the ratio of the two FI values was greater than 3:1, then in five out of six cases the coefficient of variation was greater in the mixed-FI case (the exception being exact equality of coefficient of variation). Thus Whitaker et al.'s data from rats suggests that when two standard durations are represented in reference memory, the variability of their representations is higher than when the same values are stored singly.

A notable feature of our results was the failure of superimposition of the temporal generalization gradients obtained from the short and long conditions when the "changing standard" temporal generalization method, where the standard is in force for a single trial block but changes between blocks, is employed. Any failure of superimposition is noteworthy, as non-counting-based timing behaviour in humans usually exhibits it, even when complicated procedures like the human "time-left" analogue (Wearden, 2002) are used, but in fact exactly the same result was obtained in the single previous article to report data from the changing standard method—that of Jones and Wearden (2003). In contrast, the present Experiment 3, which used a "normal" temporal generalization method like that of Wearden (1992), produced superimposition, as that method always seems to do (e.g., Droit-Volet et al., 2001; Wearden, 1992; Wearden & Towse, 1994; Wearden et al., 1997). Why does the changing standard procedure produce data that violate superimposition? We have no ready answer to this question, and any answers to it are complicated by the fact that some other temporal generalization variants *do* produce superimposition. For example, Wearden and Bray's (2001) "episodic" temporal generalization appears to dispense both with stimuli identified as "standards" and with temporal reference memory, thus differing quite radically from "normal" temporal generalization, yet this method produces superimposition, at least when conditions involving very brief stimuli are excluded. What it is about the changing standard method that causes the superimposition violation remains to be identified.

Future research with humans might explore the properties of reference memory by loading it more heavily than in the present work, for example by having more than two standard durations stored at the same time. However, storing many more than two values in reference memory might cause problems of confusion between some of them and difficulties in encoding the duration separately, as in Brown and West (1990). Another potentially informative manipulation might be to compare the behavioural effects of having two values in reference memory rather than one, but to systematically vary the difference or ratio of the two time values. For example, suppose that the two standard durations were 400 and 500 ms, compared with 400 and 800 ms. Would there be a relatively greater increase in memory variability with the putatively highly confusable values (400 and 500 ms) than with the less confusable ones (400 and 800 ms)? Given that Whitaker et al. (in press) found that the behaviour of rats on mixed-FI schedules was different in a number of ways, when the two FI values making up the mixed-FI schedule had a small ratio (less than 3:1 or 4:1 in their case) compared with larger ratios, it seems likely that this manipulation would yield interesting behavioural difference in humans also.

Overall, the present study suggests strongly that encoding two distinct temporal standards in reference memory, and having to use them both for decision about behaviour in intermixed trials, mainly increases the variability of representations in reference memory. As discussed above and elsewhere (Jones & Wearden, 2003) the reference memory component of the SET system has received little attention compared with the internal clock, particularly in work with humans. The present article takes a small step towards understanding the properties of this critical component of the SET model.

REFERENCES

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

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

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

Brown, S. W. (1998). Automaticity versus timesharing in timing and tracking dual-task performance. *Psychological Research*, *61*, 71–81.

Brown, S. W., Stubbs, D. A., & West, A. N. (1992). Attention, multiple timing, and psychophysical scaling of temporal judgments. In F. Macar et al. (Eds.), *Time, action and cognition* (pp. 129–140). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Brown, S. W., & West, A. N. (1990). Multiple timing and the allocation of attention. *Acta Psychologica*, *75*, 103–121.

Church, R. M. (1984). Properties of the internal clock. Timing and time perception. *Annals of the New York Academy of Sciences*, *423*, 566–582.

Church, R. M. (1989). Theories of timing behavior. In S. B. Klein & R. R. Mowrer (Eds.), *Contemporary learning theories: Instrumental conditioning theory and the impact of biological constraints on learning* (pp. 41–71). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Church, R. M. (1997). Timing and temporal search. In C. M. Bradshaw & E. Szabadi (Eds.), *Time and behaviour: Psychological and neurobehavioural analyses* (pp. 41–79). Amsterdam: Elsevier.

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

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., & 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.

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

Gibbon, J., & Church, R. M. (1984). Sources of variance in an information processing theory of timing. In H. L. Roitblat, T. B. Bever, & H. S. Terrace (Eds.), *Animal cognition* (pp. 465–488). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

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.

Leak, T. M., & Gibbon, J. (1995). Simultaneous timing of multiple intervals: Implications of the scalar property. *Journal of Experimental Psychology: Animal Behavior Processes*, *21*, 3–19.

Lejeune, H., & Wearden, J. H. (1991). The comparative psychology of fixed-interval responding: Some quantitative analyses. *Learning and Motivation*, *22*, 84–111.

Malapani, C., Rakitin, B., Levy, R., Meck, W. H., Deweer, B., Dubois, B., & Gibbon, J. (1998). Coupled temporal memories in Parkinson's disease: A dopamine-related dysfunction. *Journal of Cognitive Neuroscience*, *10*, 316–331.

Maricq, A. V., Roberts, S., & Church, R. M. (1981). Methamphetamine and time estimation. *Journal of Experimental Psychology: Animal Behavior Processes*, *7*, 18–30.

Meck, W. H. (1983). Selective adjustment of the speed of internal clock and memory processes. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*, 171–201.

Meck, W. H., & Church, R. C. (1987). Cholinergic modulation of the content of reference memory. *Behavioral Neuroscience*, *101*, 457–464.

Meck, W. H., Church, R. C., & Olton, D. S. (1984). Hippocampus, time and memory. *Behavioural Neuroscience*, *98*, 3–22.

Olton, D. S., Wenk, G. L., Church, R. M., & Meck, W. H. (1988). Attention and the frontal cortex as examined by simultaneous temporal processing. *Neuropsychologia*, *26*, 307–318.

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

Rakitin, B. C., Gibbon, J., Penney, T. B., Malapani, C., Hinton, S. C., & Meck, W. H. (1998). Scalar expectancy theory and peak-interval timing in humans. *Journal of Experimental Psychology: Animal Behavior Processes, 24*, 1–19.

Roberts, S. (1981). Isolation of an internal clock. *Journal of Experimental Psychology: Animal Behavior Processes*, *7*, 242–268.

Rodriguez-Girones, M. A., & Kacelnik, A. (1995). Interval bisection with and without reference memory. In M. Richelle, V. DeKeyser, G. d'Ydewalle, & A. Vandierendonck (Eds.), *Time and the dynamic control of behavior* (pp. 47–72). Liège, Belgium: Université de Liège P.A.I.

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). Travelling in time: A time-left analogue for humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*, 200–208.

Wearden, J. H., & Bray, S. (2001). Scalar timing without reference memory? Episodic temporal generalization and bisection in humans. *Quarterly Journal of Experimental Psychology*, *54B*, 289–309.

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

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

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

Whitaker, S., Lowe, C. F., & Wearden, J. H. (in press). Multiple-interval timing in rats: Performance on two-valued mixed fixed-interval schedules. *Journal of Experimental Psychology: Animal Behavior Processes*.

Original manuscript received 23 August 2002
Accepted revision received 15 March 2003

APPENDIX

Instructions given to participants in Experiments 1 and 2

In this experiment you have to decide whether different tones have the same duration as a standard tone. Each block of trials starts with three presentations of a high-pitched tone and three of a lower-pitched tone.

Before each block you will get a prompt which sometimes tells you which of the two tones is the standard. If the prompt says HIGH then you only have to attend to the high tones; if it says low it's only the LOW tone which is important. On some trials, the prompt will say EITHER, so you won't know which tone is the standard and you will have to try to remember both.

Then you receive 7 other stimuli and you have to decide whether each stimulus had the same duration as the standard and indicate your answer by pressing Y (yes it was the same duration) or N (no, it wasn't). If the prompt had been EITHER you should now use the standard with the same pitch (high or low) as the comparisons you get. The standard will change every block and you should always compare the comparison stimuli with the standard you have most recently received.

At beginning of all blocks
You will now receive 3 presentations of 2 tones which have different durations and pitches.

Alternative messages before standard presentation
1) Attend to the duration of the HIGH pitch tone. The other stimulus isn't important.
2) Attend to the duration of the LOW pitch tone. The other stimulus isn't important.
3) EITHER stimulus can be the standard. Try to remember the duration of BOTH stimuli.