ELSEVIER

Behavioural Processes 46 (1999) 25–38

BEHAVIOURAL PROCESSES

'Subjective lengthening' during repeated testing of a simple temporal discrimination

J.H. Wearden *, Richard Pilkington, Emma Carter

Department of Psychology, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Received 25 April 1997; received in revised form 27 November 1998; accepted 27 November 1998

Abstract

Three experiments investigated behavioural changes, using a temporal generalization paradigm with human subjects, resulting from repeated testing without feedback. In Experiment 1, different groups received five initial presentations of a 400 ms 500 Hz tone, or a 400 ms 14 × 14 cm blue square, identified as having a standard duration, then received blocks of testing where stimuli with durations shorter than, longer than, or equal to the standard were presented. Subjects had to judge whether each presented stimulus was the standard duration, but no feedback was given. Temporal generalization gradients (proportion of identifications of a stimulus as being the standard, plotted against stimulus duration) shifted progressively to the right during the test phase (i.e. longer stimuli tended to be identified as the standard as testing proceeded) in the visual stimulus condition. Experiment 2 used a generalization procedure to examine, with different subject groups, behavioural changes when either the duration, or the length, of a blue bar presented on the computer screen was the basis of judgement. Across trials, both length and duration could vary, but for one group only duration was relevant whereas for the other group only length was. Generalization gradients shifted systematically to the right only in duration judgements. Experiment 3 replicated the rightward shift in generalization gradients when the duration of visual stimuli was measured, and in addition used a self-rating scale derived from Thayer [Thayer, R.E., 1967. Measurement of activation through self-report. Psych. Rep. 20, 663–678.) to measure subjects' arousal. This declined systematically as testing proceeded, suggesting that the shift in temporal generalization gradient was probably caused by an arousal-induced change in internal clock speed. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Temporal discrimination; Temporal generalization; Memory for duration; Subjective shortening; Arousal measures; Humans

* Corresponding author. Tel.: +44-161-2752592; fax: +44-161-2752588.
E-mail address: wearden@psy.man.ac.uk (J.H. Wearden)

PII: S0376-6357(98)00059-X

1. Introduction

As is well known, the current leading theory of animal timing, the scalar expectancy theory of Gibbon (1977), Gibbon et al. (1984), which has also recently received considerable attention as an account of some aspects of timing in humans (e.g. Wearden, 1991; Penton-Voak et al., 1996), is a model in the internal clock tradition (Wearden and Penton-Voak, 1995, for a review of some early work using this idea). That is, it proposes that the raw material for time judgements is derived from some dedicated timing mechanism, specifically, a pacemaker-accumulator internal clock (Gibbon et al., 1984). What is less frequently emphasized, however, is that in order to derive predictions about behaviour scalar expectancy theory must invoke additional processes, in particular decisions about when or if to respond on the timing task in force and, more pertinently for present purposes, short- and longer-term memories for duration.

In spite of the central role played by memories for duration in the scalar timing system, experimental investigations of the properties of memory for duration in both animals and humans have been scarce, with the exception of work deriving from results obtained by Spetch and colleagues (e.g. Spetch and Wilkie, 1983). What was particularly interesting about these results was that they suggested that short-term memory for duration was often subject to an unusual type of forgetting, subjective shortening. This can be briefly characterised as the idea that as the short-term memory of some real-time duration t s ages, the subjective representation of t becomes progressively shorter, rather than, for example, degrading randomly (e.g. increasing or decreasing) as the memory ages. Spetch and Rusak (1992) review the literature on subjective shortening, the minutiae of which are not relevant here except for the general notion that memory for duration might have unusual properties not found in memory for most other sorts of material.

The work of Spetch and colleagues employed pigeons, but subjective shortening in short-term memory for duration was demonstrated in humans by Wearden and Ferrara (1993). The existence of subjective shortening in memory for duration raises two obvious questions. The first is whether subjective shortening is found only for duration or whether it occurs when other types of judgements are made (an issue addressed by Wearden et al., 1999). The second, which forms the focus of the present article, is whether subjective shortening of duration is found in longer-term memory. A characteristic of experiments demonstrating subjective shortening in both humans and animals has been that the memory tested has been short-term, in the sense that the memory was usable only for the duration of a single trial. For example, Spetch and Wilkie (1983) presented pigeons with a stimulus that was either 2 or 10 s long. Offset of the stimulus was followed by a retention interval, after which subjects responded on two response keys depending on the previous stimulus length, and increasing retention delays led to increasing subjective shortening in the form of an increasing tendency to report that the previously presented stimulus was the shorter one. Note that the stimulus memory is only valid for one trial at a time, as the stimulus length presented on a trial varies randomly across its two possible values from trial to trial. Although Wearden and Ferraras' 1993) studies of subjective shortening in humans used a different method, it was also true in their procedure that the memory that appeared to shorten was short-term in the sense that it was valid only for one trial, and had to be replaced by a different value on each trial of the task.

The work in the present article was motivated by what seemed an extremely simple question: What kind of forgetting would be observed if a longer-term memory for duration was tested? Here ‘longer-term’ is taken to mean a memory that is valid not only for a single trial, but for all trials of the task used. Not only was the question posed a simple one, but the means used to address it were also simple. In Experiment 1, human subjects initially received five presentations of a 400 ms duration in the form of either a tone or a visual stimulus, then were repeatedly tested using a temporal generalization paradigm (Church and Gibbon, 1982; Wearden, 1992), where stimulus durations both shorter and longer than 400 ms were presented (as well as 400 ms itself), and subjects were required to judge whether or not the presented stimulus was 400 ms long. Unlike some other studies of temporal generalization in humans (Wearden, 1991, 1992) no feedback as to performance correctness was given so as the experiment proceeded the initial stimulus duration was presumably progressively forgotten.

What kind of results might be expected? In temporal generalization experiments the usual measure of performance is a 'temporal generalization gradient', which in the case of studies with humans is the proportion of identifications of a stimulus as the standard plotted against stimulus duration. Previous studies of temporal generalization in humans (Wearden, 1991, 1992; Wearden and Towse, 1994; Wearden et al., 1997a,b) have varied stimulus durations within a range less than 1 s, as well as over multi-second values, and have also varied subject type between students and people on average about 50 years older. In all cases, the temporal generalization gradient was peaked at the standard value and was asymmetrical in the sense that stimuli longer than the standard tended to be more often confused with it than stimuli shorter by the same amount (e.g. 500 ms would be more frequently confused with a 400 ms standard than would 300 ms).

Previous research on memory for duration suggests two possible results. One is a subjective-shortening-type drift in the temporal generalization gradient as the standard duration is progressively forgotten. This would manifest itself in a leftward shift of the temporal generalization gradient; for example, if subjective shortening of the standard in memory occurs, progressively shorter and shorter stimuli would be expected to be identified as the standard. On the other hand, a second plausible possibility is 'normal' forgetting, where the memory of the standard randomly degrades, leading to a progressively flattening of the temporal generalization gradient, with some individual cases of rightward shift and others of leftward shift, but overall no systematic displacement of the generalization gradient peak. As will be seen later, neither of these results was in fact obtained.

It should be acknowledged that our procedure confounds changes in the representation of the standard that might be due to the passage of time, per se, and changes due to repeated testing without feedback (a possible analogue of non-reinforced testing in animals). However, the methodological difficulties involved in studying changes at several minutes after standard presentation without any intervening activities seem insurmountable, particularly when humans are used. In their study of short-term forgetting of duration, Wearden and Ferrara (1993) presented pairs of stimuli on each trial, with the gap between offset of the first and onset of the second ranging from 1 to 16 s. Extending this gap to several minutes would be impractical, and pose obvious difficulties of maintaining task attention. Overall, therefore, although our procedure does not provide the 'pure' measure of forgetting that a much longer-term equivalent of the Wearden and Ferrara method might provide, it is much more methodologically feasible.

2. Experiment 1: methods

2.1. Subjects

A total of 28 Manchester University first-year undergraduates participated for course credit.

2.2. Apparatus

An Opus 16X (IBM-compatible) with a standard colour monitor controlled all experimental events. The computer keyboard was used to register responses. The experimental control program was written in the MEL language (Micro-Experimental Laboratory; Psychology Software Tools), which ensured millisecond accuracy for timing of stimuli and responses.

2.3. Procedure

All subjects received a single experimental session lasting about 20 min. The procedure for the two equal-sized groups AUD (auditory stimulus) and VIS (visual stimulus) was identical except for the type of stimulus used. The procedure for the AUD group was as follows. After initial orienting instructions, subjects received five presentations of a 400 ms 500 Hz tone produced by the computer speaker. This was identified as a ‘standard duration’. Presentations of the standard were spaced by times drawn randomly from a uniform distribution running from 2000 to 3000 ms. After the tone presentations, subjects received a test phase consisting of the presentations of ten blocks of stimuli in which the standard (400 ms) duration was presented, as well as non-standard duration values. The non-standard values were 100, 200, 300, 350, 450, 500, 600, and 700 ms; that is, there were eight non-standard durations in a block. All non-standard tones had a frequency of 500 Hz. Within each block, the standard duration was presented three times, making a total of 11 trials in the block. To produce each stimulus in the block, the subject pressed the spacebar after a ‘Press spacebar for next trial’ prompt, and this response was followed by a random delay sampled from a uniform distribution running from 2000 to 3000 ms, then the stimulus presentation. Then, the display ‘Did that stimulus have the same duration as the standard? Press Y (YES) or N (NO) keys.’ was presented, and subjects made their response. No feedback as to the correctness of the response made was given during the test phase. The 11 stimuli in the block were arranged in a different random order for each block, and in testing 10 blocks were given in all. The procedure for the VIS group was identical except that the stimulus to be timed was a 14 × 14 cm light-blue square presented in the centre of the computer screen.

2.4. Results

Temporal generalization gradients in the form of proportion of YES responses (identifications of a stimulus as having the same duration as the standard, 400 ms) plotted against stimulus duration were calculated separately from blocks 1 and 2, 5 and 6, and 9 and 10 of the testing phase. Fig. 1 shows the resulting gradients, with data from the AUD group in the top panel, and data from the VIS group in the lower one.

For both the AUD and VIS groups, generalization gradients from the two blocks immediately following initial training (1/2) were (i) peaked at the standard duration (400 ms) and (ii) slightly asymmetrical in that stimuli longer than 400 ms

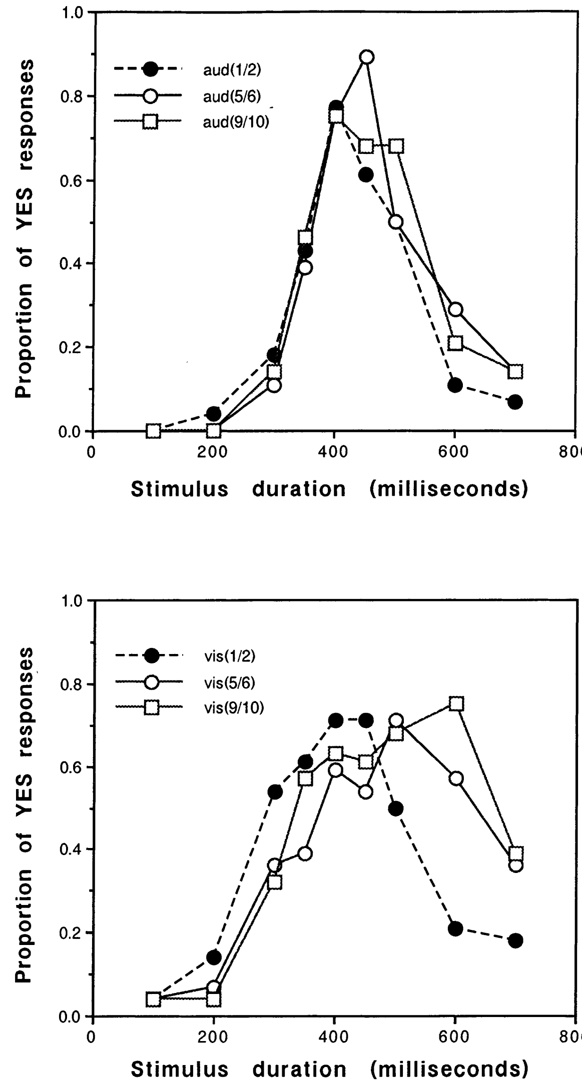

Fig. 1. Temporal generalization gradients from judgements of an auditory stimulus (top) or visual stimulus (bottom) in Experiment 1. Both panels show gradients defined as proportion of YES responses (identifications of a stimulus as having the 400 ms standard duration) plotted against stimulus duration. Filled circles connected by dashed lines show data obtained from blocks 1 and 2 of testing; open circles and squares data from blocks 5/6 and 9/10 respectively.

tended to produce more YES responses that stimuli shorter. With increasing amounts of testing (i.e. comparing blocks 5/6 and 9/10 with blocks 1/2) the temporal generalization gradients shifted rightwards, with the effect being small in the AUD group, but very marked in the VIS condition.

An overall ANOVA with group (AUD or VIS) as a between-subjects factor and stimulus duration and block (1/2, 5/6, and 9/10) as within-subjects factors produced a significant main effects of group ($F(1,26) = 7.64$, $P = 0.01$), and stimulus duration ($F(8,208) = 36.8$, $P < 0.001$), significant two-way interactions between group and stimulus duration ($F(8,208) = 2.99$, $P < 0.01$), and block by stimulus duration ($F(16,416) = 2.70$, $P < 0.001$), and a just-significant three-way interaction between group, block, and stimulus duration ($F(16,416) = 1.69$, $P = 0.045$). There were no significant effects of block or group × block.

The patterns of effects are, however, clearer when ANOVAs are conducted separately for the AUD and VIS conditions. In the AUD group, there was no effect of block ($F(2,26) = 0.75$), indicating that the overall proportion of YES responses did not increase or decrease over blocks. The effect of stimulus duration was significant ($F(8,104) = 27.92$, $P < 0.001$), indicating that different stimulus durations produced, overall, different proportions of YES responses. This effect is obvious on inspection of Fig. 1 and merely shows that subjects, unsurprisingly, were sensitive to stimulus duration. The interaction between stimulus duration and block, which might indicate a shift in the shape of the stimulus generalization gradient over blocks, did not, however, reach significance ($F(16,208) = 1.32$, $P = 0.18$), in spite of the suggestion derived from inspection of the data in Fig. 1 that the temporal generalization gradient shifted to the right in the AUD group as forgetting proceeded.

For the VIS group, there was no overall effect of block ($F(2,26) = 1.07$, $P = 0.36$), but there was a significant effect of stimulus duration ($F(8,104) = 13.48$, $P < 0.001$), and a significant block by stimulus duration interaction ($F(16,208) = 2.74$, $P = 0.001$). Thus, in the VIS group, not only were subjects highly sensitive to stimulus duration, but their generalization gradients changed significantly between blocks.

Another way of examining changes in the shape of temporal generalization gradients over blocks of testing is to construct an 'asymmetry score'. We defined this as the average proportion of YES responses to stimuli longer than 400 ms minus the average proportion of YES responses to stimuli shorter. Here a positive value indicates gradient asymmetry, with more YES responses to stimuli longer than the standard than shorter. How did this value change over blocks? Average values for the AUD group were 0.16, 0.28, and 0.33, for blocks 1/2, 5/6, and 9/10, respectively. Thus all values were positive (indicating asymmetry), and the asymmetry increased (indicating an increasing rightward skew the temporal generalization gradient) over blocks. However, ANOVA revealed no significant effect of block ($F(2,26) = 1.74$, $P = 0.19$). For the VIS group, on the other hand, the respective mean asymmetry scores were 0.07, 0.33, and 0.37 (all positive and increasing with blocks), and for this group a significant increase did occur ($F(2,26) = 6.25$, $P < 0.01$). This supports statistically the impression gained from inspection of Fig. 1 that for the VIS group the temporal generalization gradient not only changed over blocks, but shifted systematically rightwards.

2.5. Discussion

In Experiment 1, temporal generalization gradients shifted progressively to the right (that is, progressively longer and longer stimuli were identified as the standard) as testing without feedback proceeded, an effect which was highly significant in the visual stimulus condition. Inspection of data from the auditory condition suggested that a weaker, albeit statistically non-significant, effect was also present there; at the very least, data from the AUD group did not exhibit a leftward shift in gradient.

The rightward gradient shift was unexpected, and contrary to both of the predictions derived from previous work on forgetting (subjective shortening: leftward shift of the gradient; random forgetting: flattening of the gradient but no systematic shift). At first sight, our results suggest that the standard duration remembered in our paradigm exhibited a kind of 'subjective lengthening' as forgetting proceeded so, for example, the remembered standard duration became progressively longer in memory, thus stimuli with longer and longer durations were identified with the standard as it is forgotten. Such an effect would indeed produce the pattern of results we observed but, as will be seen later, it is not the only possibility.

Given the unusual results of our Experiment 1, we decided to replicate them in Experiment 2, but Experiment 2 in addition addressed another question. Suppose that humans behave as if the stimulus duration remembered lengthened in memory as the memory aged: Is this effect unique to duration, or can it occur in judgements of other sorts of material? Strictly speaking, this question is probably impossible to answer, as 'uniqueness' cannot be demonstrated without testing all conceivable types of stimulus judgement, and finding 'subjective lengthening' only in duration, which obviously cannot be done. Experiment 2 did, however, employ a method which partially addressed this issue.

The stimuli used in Experiment 1 differed from one trial to another only in duration, and other stimulus dimensions (for example, pitch and intensity of tone, colour or area of the square) were both invariant and irrelevant. But suppose that stimuli are used which vary between trials along two dimensions, both of which could potentially demonstrate 'directional forgetting' effects like subjective shortening or lengthening. In Experiment 2, one of these dimensions was duration and the other was length, so in different groups we can examine the changes contingent on repeated testing of a standard duration, or of a standard length.

Experiment 2 employed a coloured bar presented on the computer screen. The standard stimulus was always the same: the bar was presented for 400 ms, and had a fixed length (200 pixel units). Following presentation of the standard, test blocks of stimuli were presented which involved bars which had greater or lesser durations than the standard, as well as longer or shorter lengths and, across blocks, length and duration were not associated. Two groups of subjects were used, and both made judgements of stimuli which were physically identical (the coloured bar which differed between trials in both length and duration). The difference was that for one group (DUR) the duration of the stimulus was the basis for the judgement of whether or not it was the standard, and for the other group (LEN) it was the length of the stimulus that was the basis of the judgement. Thus, for the DUR group length was irrelevant, and for the LEN group duration was irrelevant, but both the standard stimulus initially given, and all stimuli subsequently presented, were physically identical for the two groups.

Of course, the length and duration judgements made of the same stimuli are not completely comparable: for one thing, a length judgement can be made immediately after stimulus onset, whereas the duration judgement can only be produced after offset of the stimulus; for another, the different discriminations may differ in difficulty. However, the point of including the length task was to examine if simple repeated testing of a discrimination without feedback by itself produced a rightward shift in the generalization gradient, regardless of the stimulus dimension being judged.

What would performance look like during the testing blocks given without feedback? From the results of Experiment 1 we would expected an apparent 'subjective lengthening' effect in the DUR group, but in the LEN group there are obviously other possibilities. One of these is that length, like duration, might also be subject to some sort of 'directional forgetting' (thus the generalization gradients for length might shift progressively leftwards (subjective shortening) or rightwards (subjective lengthening) as forgetting proceeded). An alternative possibility is that of 'normal' random forgetting with length, that is, a progressive flattening of the generalization gradient without any systematic shift in peak. This latter result would suggest that directional forgetting was a property of duration memory rather than memory for other stimulus dimensions.

3. Experiment 2: methods

3.1. Subjects

A total of 30 Manchester University undergraduate students participated.

3.2. Apparatus

A Hyundai TSS8C (IBM-compatible) computer controlled all experimental events. Other details were as Experiment 1.

3.3. Procedure

All subjects received a single experimental session lasting about 20 min. The procedure for both equal-sized groups was identical, as were the standard stimuli, and all other stimuli presented for judgement in the experiment. The only difference between the two groups was the stimulus dimension that was relevant to the judgement made. For one group (DUR) it was the duration of the standard which had to be remembered, and duration was the basis of all judgements to be made (thus changes in length between trials were irrelevant). For the other group (LEN) it was the length of the standard and all subsequent stimuli that was the basis of judgement, and variation in duration between trials was irrelevant. Subjects received instructions to this effect before the experimental procedure started. The standard stimulus (a light-blue bar or thick line, 200 pixel units long and five wide (about 11.2×0.28 cm), 400 ms in duration) was initially presented five times, separated by random interpresentation intervals drawn from a uniform distribution running from 2000 to 3000 ms. Following this subjects received a testing phase, and during this were required to respond YES if they judged that the stimulus just presented had the same duration (DUR group) or the same length (LEN group) as the standard, and NO otherwise. No feedback was given. There were 11 stimuli in each test block, as in Experiment 1. The durations used were 100, 200, 300, 350, 400 (standard), 450, 500, 600, and 700 ms, and the lengths (in pixel units) were 140, 155, 170, 185, 200 (standard), 215, 230, 245, 260 (range from 7.84 to 15.46 cm). In each block the standard length or duration was presented three times, the other eight non-standard values once each. Within a block, when the standard was presented it always had the same length for the DUR group, and the same duration for the LEN group, but the irrelevant length and duration were not necessarily the same as for the standard stimulus itself, and the irrelevant value associated with the standard varied between blocks. For the non-standard stimuli, the relation between duration and length was systematically varied between blocks, so that duration and length were not associated overall. In summary, both the duration and length of stimuli varied between trials and blocks, but for the DUR group duration was relevant and the basis of judgement, whereas for the LEN group, length alone was relevant.

During testing blocks, the subject pressed the spacebar in response to a ‘Press spacebar for next trial prompt’, and this was followed after a random delay drawn from a uniform distribution running from 2000 to 3000 ms by the stimulus. Then, the subject was prompted to judge whether or not the stimulus had the standard value (duration or length, depending on group), but no feedback was ever delivered after a YES or NO response. The 11 stimuli within the block were presented in a random order which differed between subjects. Nine blocks of stimuli to be judged were presented.

3.4. Results

In Experiment 2, there were nine blocks of testing without feedback after the initial stimulus presentations. Generalization gradients were constructed from blocks 1/2, 4/5, and 8/9. To facilitate visual comparison between the groups, the generalization gradients were plotted as proportion of YES responses against stimulus number, where stimulus 5 was the standard (duration, 400 ms; length, 200 pixels), stimuli 1–4 were the stimuli of lower value (duration or length) than the standard; stimuli 6–9 those of higher value. Fig. 2 shows the results, with gradients from the DUR group (duration discrimination) in the top panel and the LEN group (length discrimination) in the bottom panel

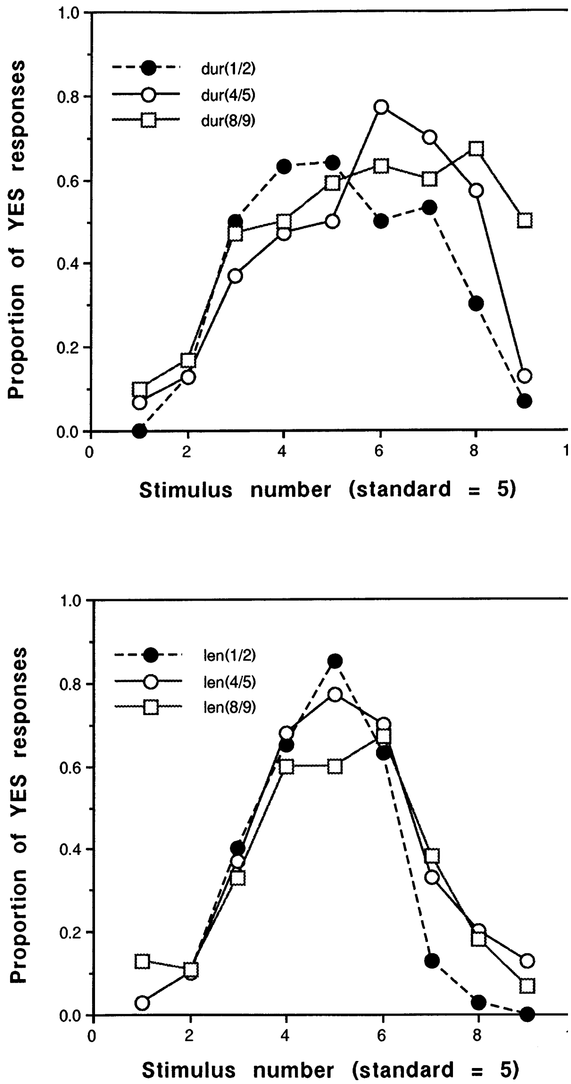

Fig. 2. Generalization gradients from Experiment 2. (top) DUR group (duration relevant; length irrelevant). The proportions of identifications of a presented stimulus as the standard (i.e. being 400 ms long) are plotted against stimulus number. 1–4 indicate stimuli with shorter durations than the standard, 5 is the standard, 6–9 stimuli with durations longer than the standard. (bottom) LEN group (length relevant; duration irrelevant).The proportions of identifications of a stimulus as having the standard (200 pixel unit) duration is plotted against stimulus number. 1–4: stimuli shorter than the standard, 5 is the standard 6–9 stimuli longer than the standard. Both panels show data separately from testing blocks 1/2 (•), and blocks 4/5 (○) and 8/9 (□).

Inspection of the generalization gradients from the two groups suggests that, in spite of the fact that the stimuli judged were physically identical, repeated testing without feedback of the duration dimension (DUR) produced a different pattern of behavioural change from that obtained in the length discrimination (LEN). In the DUR group, the generalization gradient appeared to shift systematically rightwards as testing proceeded, whereas for the LEN group there was no such clear systematic shift in gradient peak.

ANOVA of the data from the two groups conducted separately confirmed the above suggestions. For the DUR group, there was a significant effect of stimulus duration ($F(8,112) = 10.68$, $P < 0.001$), and a significant block by duration interaction ($F(16,224) = 2.66$, $P = 0.001$), whereas the effect of block just failed to reach significance ($F(2,28) = 3.17$, $P = 0.06$). In this case, as in Experiment 1, it is the stimulus duration by block interaction that confirms the change in generalization gradient with repeated testing. The asymmetry measure used in Experiment 1 was again calculated. Mean values for blocks 1/2, 4/5, and 8/9 were 0.09, 0.33, and 0.39 respectively. Thus all mean asymmetry scores were positive (indicating right-skewed asymmetry in the generalization gradient), and values increased with increasing amount of testing. ANOVA of asymmetry values showed a significant effect of block ($F(2,28) = 4.63$, $P = 0.02$) confirming that the shift in the generalization gradient with blocks was an increasing rightward skew (i.e. more YES responses at longer durations).

In the LEN group, there was a significant effect of stimulus length ($F(8,112) = 16.13$, $P < 0.001$), but neither the effect of block ($F(2,28) = 1.87$, $P = 0.17$) nor the block by stimulus length interaction ($F(16,224) = 1.35$, $P = 0.17$) were significant. Mean asymmetry score values for blocks 1/2, 4/5, and 8/9, respectively, were −0.10, 0.03, and 0.05, and ANOVA revealed that these did not change significantly over blocks ($F(2,28) = 1.85$, $P = 0.17$).

3.5. Discussion

The results of Experiment 2 replicated the main finding of Experiment 1: When humans per-formed temporal generalization over a 100 or more trials without feedback (about 20 min), their generalization gradients shifted progressively to the right with increasing amounts of testing, consistent with the idea of subjective lengthening of long-term memory for duration of the standards. On the other hand, when the basis of judgement was the length of the same stimuli used in the duration task, the pattern of change with repeated testing was quite different, with no systematic shift in the peak location of the generalization gradient. The different patterns of change over testing blocks when the target stimulus dimension was either duration of length were observed even though the same physical stimuli were being judged; thus the ‘directional forgetting’ observable in the duration perception appeared to be due to the type of processing the stimulus had received, rather than what the physical stimulus actually was. This result parallels those obtained in Wearden et al. (1999), who found that the length and duration of a bar-like stimulus had different forgetting characteristics in a Wearden and Ferrara (1993) short-term memory paradigm.

Why does our ‘subjective lengthening’ effect occur? There are obviously two different types of explanation. One, a true ‘memory account’, would propose that for some reason, the long-term representation of the standard duration changes in the direction of becoming apparently longer as the testing phase proceeds. Another explanation might be that the memory of the standard duration remains constant during the testing phase, but that some other sort of change is responsible for the effects observed. But what could this change be?

One possibility is that pacemaker rate is determined by the arousal level of the subject, and that the rather uninvolving judgement task we used leads to a decrease in arousal as the experiment proceeds, and produces the apparent ‘subjective lengthening’ effect by this means. An arousal-based explanation of the rightward gradient shift is as follows. Suppose subjects initially encode the standard duration as n ‘ticks’ from a pacemaker which runs at some average constant rate, r. As the testing phase proceeds, arousal falls, so r decreases. In effect, this means that longer and longer stimuli will be needed to provide the n ticks that the subject has encoded as the standard duration, so longer and longer stimuli are chosen as the testing phase proceeds. This explanation assumes that the representation of the standard duration remains constant during testing, so the results are not due to any change in memory of n, but instead due to a change in pacemaker speed.

Various authors have provided evidence that the rate of underlying pacemakers possessed by both humans and animals varies with arousal. Treisman et al. (1990), Boltz (1994), Penton-Voak et al. (1996) all showed effects on time perception in humans consistent with this idea, see also Wearden and Penton-Voak (1995). Furthermore, Fetterman and Killeen (1995) provided data from one of the few relevant animal experiments, when they showed that increasing or decreasing the rate of reinforcement in a categorical timing task shifted timing behaviour in a way consistent with their idea that increasing reinforcement rate increased pacemaker speed whilst decreasing reinforcement rate slowed the pacemaker, by respectively increasing and decreasing the animal’s arousal.

Can the arousal-based explanation of ‘subjective lengthening’ be tested? One way would be to use a standard self-rating scale of arousal, which subjects could complete at different times during the testing phase. Not only should we see a decrease in arousal manifested in the ratings obtained on this scale, so providing an independent measure of arousal changes, but the rightward shift in the temporal generalization gradient should parallel these changes in self-rated arousal.

Experiment 3 closely followed the method used for the VIS group of Experiment 1. Subjects received initial presentations of a visual standard 400 ms long, then a testing session in which stimuli of various durations, including the standard duration, were presented without feedback. At various times during the testing phase, subjects filled in a brief questionnaire derived from Thayer (1967), which gave measures of ‘activation’ (indicating a high arousal state), and ‘deactivation’ (indicating lower arousal).

4. Experiment 3: methods

4.1. Subjects

A total of 23 Manchester University undergraduates participated for course credit.

4.2. Apparatus

The temporal generalization procedure was arranged using the same apparatus as Experiment 2. In addition, a cardboard box with a slit in its lid was used as a 'post-box' for the questionnaire.

4.3. Procedure

All subjects received a single experimental session. This began with 5 presentations of the standard stimulus (a 14×14 cm light blue square in the centre of the computer screen) for 400 ms. The presentations were spaced by randomly chosen intervals running between 1000 and 2000 ms. Five blocks of testing then followed. Within each block, the standard duration was presented three times, and there were eight non-standard durations; 100, 200, 300, 350, 450, 500, 600, and 700 ms, each presented once. The 11 stimuli were arranged in a random order which varied between blocks. To produce each stimulus, the subject pressed the spacebar in response to a 'Press spacebar for next trial' prompt, and the stimulus was delivered between 2000 and 3000 ms later. After stimulus offset, the display 'Did that stimulus have the standard length? Press Y or N keys' appeared, and the subject made the response. No feedback as to the correctness of the response was given during the testing phase.

At the end of each block of 11 stimuli, the experiment was halted and the subject filled in the questionnaire shown in Fig. 3. After completion, the numbered questionnaire sheet was deposited in the 'post-box' (a precaution used to prevent subjects altering their questionnaires retrospectively), and the next block of testing was given.

The self-report scale was a condensed version of Thayer's (Thayer, 1967) activation–deactivation adjective checklist. In its original form, the scale consisted of 28 activation adjectives grouped, on the basis of results from factor-analytic studies, into four activation groupings: general activation, high activation, general deactivation, and deactivation sleep. To produce the scale shown in Fig. 3, two adjectives were randomly selected from each of the four groupings. These were lively, energetic (general activation); jittery, stirred-up (high activation); calm, relaxed (general deactivation); drowsy, sleepy (deactivation sleep). Note that responses on the scales are arranged from 1 to 4, ranging from 'definitely feel' to 'definitely do not feel', so higher scores mean less of the quality measured by the adjective.

4.4. Results

Fig. 4 shows measures derived from the temporal generalization gradients over the five blocks of testing without feedback (top), and scores derived from the Thayer scale (bottom).

In the top panel, the three measures shown are (a) the average proportion of yes responses (identifications of a stimulus as having the standard, 400 ms, duration) for stimuli that were longer than the standard (LONG), (b) the same measure for stimuli shorter than the standard (SHORT), and (c) an asymmetry score LONG − SHORT (ASYM). Obviously, LONG scores tended to in-

When the prompt to fill in the ratings scale appears on the computer, please circle the number which most accurately applies to your present state against each of the following adjectives.

	Definitely Feel	Feel slightly	Cannot decide	Definitely do not feel
ENERGETIC	1	2	3	4
DROWSY	1	2	3	4
CALM	1	2	3	4
JITTERY	1	2	3	4
SLEEPY	1	2	3	4
LIVELY	1	2	3	4
STIRRED-UP	1	2	3	4
RELAXED	1	2	3	4

Once you have completed this scale, please post it into the cardboard "post-box" to the left of the computer.

Fig. 3. Rating scale used at the end of each block of testing during Experiment 3.

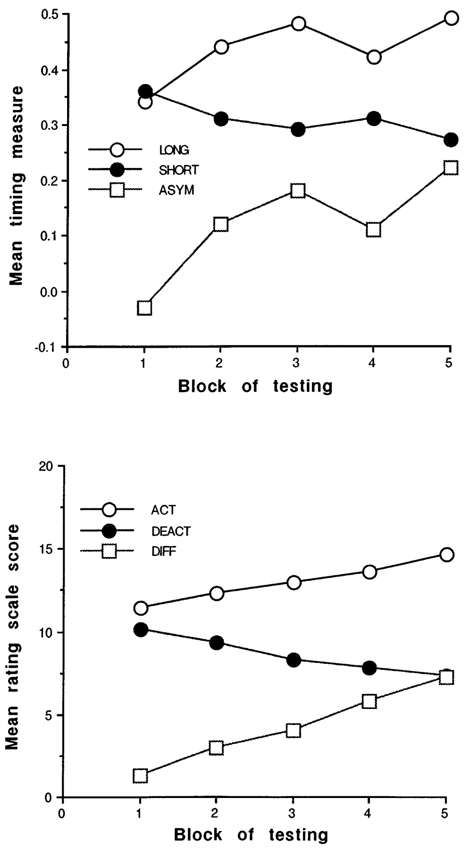

Fig. 4. Top: Timing measures derived from temporal generalization gradients obtained in Experiment 3. Measures shown are mean proportion of responses to stimulus durations longer than the standard (LONG: ○) and shorter than the standard (SHORT: ●), plotted against number of test blocks. Also shown is the measure of generalization gradient asymmetry (ASYM = LONG − SHORT: □). Bottom: Measures derived from Thayer (1967) rating scale plotted against number of test blocks. Measures of activation (ACT: ○) and deactivation (DEACT: ●) are shown separately. For both measures, higher scale scores mean less of the attribute shown (see Fig. 3). Also shown is the difference ACT–DEACT (DIFF: □).

crease with increasing test blocks, and the SHORT scores to decrease, indicating an increasing rightward shift, and this was manifested in the increasing asymmetry score. The asymmetry scores significantly increased over blocks of testing ($F(4,88) = 3.07$, $P < 0.05$), and the increase in the LONG measure was also significant ($F(4,88) = 2.75$, $P < 0.05$), but the decrease in the SHORT measure did not reach significance ($F(4,88) = 1.26$), even though on the fifth test block, the SHORT measure was significantly smaller than the LONG one ($P = 0.01$ by Wilcoxon test), whereas there was no significant difference on the first test block.

The bottom panel shows measures derived from the Thayer rating scales, with activation scores (ACT) and deactivation scores (DEACT) shown separately, as well as the difference between them (activation–deactivation: DIFF). Recall that higher scores mean less of the factor measured. Inspection of the bottom panel of Fig. 4 suggests that activation decreased (i.e. ACT scores increased), and deactivation increased with increasing number of test blocks. Consequently the difference score also increased with increasing blocks. All these changes were significant: DEACT, $F(4,88) = 31.31$, $P < 0.001$; ACT, $F(4,88) = 34.85$, $P < 0.001$; DIFF: $F(4,88) = 64.51$, $P < 0.001$).

4.5. Discussion

The shift in temporal generalization gradient with increasing amounts of testing without feedback obtained in Experiment 3 closely paralleled that obtained in Experiments 1 and 2. Generalization gradients shifted significantly to the right as testing proceeded and this was manifested in Fig. 4 by increasing LONG and ASYM scores. The contribution of Experiment 3 was the addition of independent measures of arousal, obtained from a modified Thayer scale (Thayer, 1967). The scale scores proved highly sensitive to repeated testing, with deactivation systematically increasing during the testing session and activation decreasing; in other words, arousal systematically fell as testing proceeded. Comparison of the top and bottom panels of Fig. 4 shows that the measures of timing behaviour from the temporal generalization gradients and the measures of arousal changed in parallel; as arousal decreased, the temporal generalization gradient shifted to the right. This result strongly suggests that the arousal-based interpretation of the 'subjective lengthening' effects shown in all three experiments is the correct one. That is, subjects encode a certain number of 'ticks' from an internal clock to represent the standard duration, the clock ticks more slowly as arousal falls, so longer and longer stimuli need to be chosen to obtain the required number of ticks.

5. General discussion

The three experiments of the present article together report what appears to be a newly discovered phenomenon: When humans make repeated judgements of the duration of a stimulus without feedback for a period of minutes, they behave as if the duration representation they possess systematically lengthens. Experiment 1 showed that this effect occurred for the duration of a visual stimulus, and Experiment 2 showed that when the basis of judgements of physically identical stimuli was duration or length in different groups, an apparent 'subjective lengthening' effect occurred only for duration representations. Experiment 3 replicated the results of the previous experiments, and linked the shift in temporal generalization gradients to changes in arousal measured by a self-rating scale.

We initially address two issues here. Firstly, is the phenomenon that we report present unique to our data, and where would one look for it if it were not? Secondly, what are the methodological implications of our results for studies of timing?

There are very few studies which have explicitly studied memory for duration in any context, and ours appears to be the first to have asked the question of how performance on a simple temporal discrimination changes over a period of minutes of testing without feedback. However, our results may have parallels, or at least potential parallels, in other studies. One of these comes from the extinction of temporal discriminations formed by animals. The similarity between the changes that we found were induced by repeated testing without feedback and the extinction that occurs when reinforcement is withdrawn in animal experiments needs to be treated with caution, and the ready equation of feedback/no feedback in experiments with humans and reinforcement/extinction in animal experiments can be risky (Wearden, 1988). For example, in our experiment, if feedback is considered as 'reinforcement' then the responses which measure the duration or length discrimination (YES or NO) were never reinforced, as they were not made during the initial training phase, and were not followed by feedback during the testing phase. Another point of difference is that extinction in animal experiments usually decreases the overall rate at which the previously reinforced response is made, whereas in our experiments the total number of YES responses did not change. Nevertheless, the parallel might be maintained with these cautionary notes in mind.

Unfortunately, effects of extinction of temporal discriminations in some commonly used animal timing tasks may not provide much relevant data. For example, on a fixed-interval (FI) schedule of food, the first response occurring t s after the previous food delivery is itself reinforced, and this schedule generates the well-known pattern of temporally regulated behaviour, the fixed-interval 'scallop' (Ferster and Skinner, 1957; Dews, 1978). Since delivery of food initiates the period to be timed, however, performance in extinction cannot be interpreted so far as its effects on temporal regulation are concerned. If a temporal differentiation schedule like differential-reinforcement-of-low-rate, where responses spaced by more than t s from the previous response are reinforced, is subject to extinction, although the times between responses can be measured, extinction will cause response rate to fall, so complicating the interpretation of any shift in the temporal distribution of inter-response times (see Harzem, 1969, for discussion).

However, Fetterman and Killeen (1995) reported something similar to the effects reported in the present experiments from their categorical timing procedure with pigeons. Here, three response alternatives were available on a trial, but only one of these produced the reinforcer on any particular trial, and which one potentially 'paid off' depended on the elapsed time in the trial. So, for example, one alternative would pay off (if it did on that trial) at 4 s, another at 8 s, another at 16 s. Although none of Fetterman and Killeen's conditions involved extinction, some involved a decrease in the probability of reinforcement, so the pigeons could experience a number of intervals without reinforcement. This shifted generalization-gradient-like measures of responding rightwards (Fetterman and Killeen, 1995, Fig. 10), consistent with results we report here. A more recent experiment by Killeen et al. (1999) did in fact examine behavioural changes produced by a period of extinction on a timing task and attributed the effects observed to an extinction-induced decrease in the rate of an underlying pacemaker, although extinction appeared to produce other effects as well. Overall, therefore, it appears that what few data from animal studies there are generally support our explanation in terms of changes in pacemaker speed contingent upon extinction-like manipulations.

The second issue is the methodological implications of our results for studies of timing. Training humans to identify or produce certain durations, then later testing them without feedback, is a potentially attractive procedure, and may enable the replication with humans of classic timing experiments previously conducted with animals such as studies of the effects of drugs (e.g. Maricq et al., 1981; Meck, 1983). For example, suppose subjects are initially trained in a non-drug state to learn duration, *t*, then are subsequently tested after drug administration. The basic idea here might be that the pacemaker of an putative internal clock runs faster or slower in the drug state (e.g. Maricq et al., 1981), so changes in behaviour (such as rightward or leftward shifts of temporal generalization gradients) might indicate a change in underlying pacemaker rate. However, if feedback is provided during the drug phase, subjects may rapidly 're-calibrate' to take into account the faster or slower pacemaker speed, so any drug effects would quickly disappear and, indeed, may persist only for the very first trial before feedback. On the other hand, if testing without feedback were conducted, such re-calibration would be impossible. However, our results suggest that, irrespective of effects of drug in an experiment like the one outlined, there may be systematic directional shifts in the measures of timed behaviour contingent on repeated testing without feedback, so interpretation of no-drug/drug shifts like the one described become complicated. A full range of control conditions (no-drug/no-drug and drug/drug as well as the conditions where drug state differs between training and testing) would clarify the effect of the drug, but the possibility of behavioural changes like the ones reported in the present article needs to be borne in mind in all studies where two conditions A and B are compared within subjects, when the first involves training in some state A, with subsequent testing without feedback in state B.

Finally, we return briefly to the issue of why our 'subjective lengthening' effect occurs at all. Results from our Experiment 3 clearly implicate an arousal-related change in timing processes as being the cause of the effect, rather than any change in the memory representation of the standard duration. It is, of course, possible that the arousal change is epiphenomenal and plays no role in determining timed behaviour, and that in fact an additional memory change is occurring. However, this explanation seems unacceptable without further evidence, on grounds of parsimony, even though it cannot be logically excluded. Even if we suppose that arousal changes are at the root of the observed changes in temporal generalization gradients, our proposed causal mechanism (a change in pacemaker speed) is not the only possibility. Another is that the fall in arousal causes attentional changes, so that less and less attention is paid to presented stimuli as the testing phase proceeds, and this itself results in a change in their time representations, for example by the mechanism of 'missed pulses'. Macar et al. (1994) have shown that diverting attention away from the duration of stimuli and towards other stimulus characteristics can decrease their perceived duration, and it is possible that decreased 'attention' resulting from falling arousal produces this effect in our studies. It is, however, at present unclear how this explanation could be distinguished from one based on arousal-induced changes in pacemaker speed and, furthermore, the missed-pulses explanation, like the one we advance, also assumes that the memory representation of the standard does not change during the experiment.

Overall, our three experiments taken together illustrate a robust apparent 'subjective lengthening' effect, which seems important for future studies of timing in humans where standard durations are learned and subsequently tested without feedback. Our data also further suggest a link between arousal and the speed of the pacemaker of a hypothesized internal clock, and thus join a considerable body of diverse results supporting this general idea (e.g. Maricq et al., 1981; Treisman et al., 1990; Fetterman and Killeen, 1995; Wearden and Penton-Voak, 1995; Penton-Voak et al., 1996).

Acknowledgements

We are grateful to members of a 2nd. year 'TP' group for running the subjects in Experiment 2 for us.

References

Boltz, M.G., 1994. Changes in internal tempo and effects on the learning and remembering of event durations. J. Exp. Psychol. Learn. Mem. Cognit. 20, 1154–1171.

Church, R.M., Gibbon, J., 1982. Temporal generalization. J. Exp. Psychol. Anim. Behav. Proc. 8, 165–186.

Dews, P.B., 1978. Studies on responding under fixed-interval schedules of reinforcement: II. The scalloped pattern of the cumulative record. J. Exp. Anal. Behav. 29, 67–75.

Ferster, C.B., Skinner, B.F., 1957. Schedules of Reinforcement. Appleton-Century-Crofts, New York.

Fetterman, J.G., Killeen, P.R., 1995. Categorical scaling of time: implications for clock-counter models. J. Exp. Psychol. Anim. Behav. Proc. 21, 43–63.

Gibbon, J., Church, R.M., Meck, W., 1984. Scalar timing in memory. In: Gibbon, J., Allan, L. (Eds.), Annals of the New York Academy of Sciences, Timing and Time Perception, vol. 423. New York Academy of Sciences, New York, pp. 52–77.

Gibbon, J., 1977. Scalar expectancy theory and Weber's law in animal timing. Psych. Rev. 84, 279–325.

Harzem, P., 1969. Temporal discrimination. In: Gilbert, R.M., Sutherland, N.S. (Eds.), Animal Discrimination Learning. Academic Press, London, pp. 299–334.

Killeen, P., Bizo, L.A., Hall, S., 1999. A clock not wound runs down. Behav. Proc. 45, 129–139.

Macar, F., Grondin, S., Casini, L., 1994. Controlled attention sharing influences time estimation. Mem. Cognit. 22, 673–686.

Maricq, A.V., Roberts, S., Church, R.M., 1981. Metamphetamine and time estimation. J. Exp. Psychol. Anim. Behav. Proc. 7, 18–30.

Meck, W.H., 1983. Selective adjustment of the speed of internal clock and memory processes. J. Exp. Psychol. Anim. Behav. Proc. 9, 171–201.

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. J. Exp. Psychol. Anim. Behav. Proc. 22, 307–320.

Spetch, M.L., Rusak, B., 1992. Time present and time past. In: Honig, W.K., Fetterman, J.G. (Eds.), Cognitive Aspects of Stimulus Control. Erlbaum, Hillsdale, NJ, pp. 47–67.

Spetch, M.L., Wilkie, D.M., 1983. Subjective shortening: a model of pigeons' memory for event duration. J. Exp. Psychol. Anim. Behav. Proc. 9, 14–30.

Thayer, R.E., 1967. Measurement of activation through self-report. Psych. Rep. 20, 663–678.

Treisman, M., Faulkner, A., Naish, P.L.N., Brogan, D., 1990. The internal clock: evidence for a temporal oscillator underlying time perception with some estimates of its characteristic frequency. Perception 19, 705–748.

Wearden, J.H., Ferrara, A., 1993. Subjective shortening in humans' memory for stimulus duration. Q. J. Exp. Psychol. 46B, 163–186.

Wearden, J.H., Penton-Voak, I.S., 1995. Feeling the heat: body temperature and the rate of subjective time, revisited. Q. J. Exp. Psychol. 48B, 129–141.

Wearden, J.H., Towse, J.N., 1994. Temporal generalization in humans: three further studies. Behav. Proc. 32, 247–264.

Wearden, J.H., Denovan, L., Fakhri, M., Haworth, R., 1997a. Scalar timing in temporal generalization in humans with longer stimulus durations. J. Exp. Psychol. Anim. Behav. Proc. 23, 502–511.

Wearden, J.H., Wearden, A.J., Rabbitt, P.M.A., 1997b. Age and IQ effects on stimulus and response timing. J. Exp. Psychol. Hum. Percep. Perf. 23, 962–979.

Wearden, J.H., Parry, A., Stamp, L., 1999. Is subjective shortening in human memory unique to time representations? Q. J. Exp. Psychol. (submitted).

Wearden, J.H., 1988. Some neglected problems in the analysis of human operant behavior. In: Davey, G., Cullen, C. (Eds.), Human Operant Conditioning and Behavior Modification. Wiley, Chichester, pp. 197–224.

Wearden, J.H., 1991. Do humans possess an internal clock with scalar timing properties? Learn. Motiv. 22, 59–83.

Wearden, J.H., 1992. Temporal generalization in humans. J. Exp. Psychol. Anim. Behav. Proc. 18, 134–144.