

Do Differences in Event Descriptions Cause Differences in Duration Estimates?

ALICE C. I. PEDERSEN¹ and DANIEL B. WRIGHT^{2*}

¹*Institute of Education, UK*

²*University of Sussex, UK*

SUMMARY

The relationship between the way in which people describe an event and people's estimates of the duration of the event is investigated in three studies. People are told to use different writing styles designed to produce different characteristics. For example, a 'tabloid' condition was designed to produce words with higher implied action. Across all three studies, differences among the event descriptions only produced small differences in the duration estimates. These results question the direct causal relation between language use and duration estimates. We discuss these findings in relation to memory reconstruction and eyewitness testimony. Copyright © 2002 John Wiley & Sons, Ltd.

At 9.02 am on 19 April 1995, 168 people died and over 600 people were injured when approximately one third of the Alfred P. Murray Federal Building in Oklahoma City was blown up. Gulf war veteran Timothy McVeigh was arrested 77 miles away at 10.20 am. He was later convicted of this crime (*United States v. McVeigh*, 1997). Much of the debate about this case was whether McVeigh worked with a mysterious 'John Doe 2' (Memon and Wright, 1999; Schacter, 2001; Wright *et al.*, 2000). Consider the testimony of one witness, Germaine Johnston. She worked in the Murray Building and was inside when the bomb exploded. As she left the building she saw a car with two people standing by it. When she saw the picture of McVeigh on the television she recognized him as one of the people she saw. The other man she remembered as shorter and with a darker complexion, which matches other descriptions of the mysterious 'John Doe 2'. Johnston said that this was at about 9.30 am. In order for this to be McVeigh, and therefore lead credence to the idea that there was a second man involved in planting the bomb, McVeigh would have had to leave immediately after the encounter with Johnston and drive at an average of over 90 mph to where he was arrested. This is unlikely. However, this assumes that Johnston accurately estimated that it was 28 minutes from when the bomb exploded to when she saw the two men.

The ability of people to make accurate retrospective duration estimates¹ can be directly relevant to many court cases, as the example above demonstrates. It can also have an indirect influence because duration of exposure is related to accuracy of testimony (Yarmey and Matthys, 1990). Loftus and colleagues (1987, p. 4) point out that this is

*Correspondence to: Daniel B. Wright, Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton, BN1 9QH, UK. E-mail: DanW@cogs.susx.ac.uk

¹In this paper we discuss only retrospective duration estimates, not duration estimates while experiencing the event or estimating how long a future event will last. For the remainder of the paper we just use 'duration estimate' to refer to 'retrospective duration estimate'. For a broad overview on duration estimation see Zakay and Block (1997).

'common sense as well as a psychological finding, that the longer a witness has to look at something the more accurate the memory'. Thus, the credibility of a witness may be strengthened when he or she reports having observed the crime for a relatively long period. In fact, in the United Kingdom judges (Judicial Studies Board, updated August 1999) are instructed to tell jurors to 'examine carefully the circumstances in which the identification by each witness was made. How long did he have the person he says was the defendant under observation?' (<http://www.jsboard.co.uk>). Accordingly, the general accuracy of eyewitnesses' estimates of duration is important (Yarmey, 2000). Both over- and under-estimation can have serious consequences (Burt, 1999).

Our interest is the relationship between the words used to describe an event and the estimated duration of the event. In particular, after seeing a crime an eyewitness may be asked by different people to describe the event and may adopt different communication styles. We ask how event descriptions of different lengths and styles may influence duration estimation. After describing the way in which duration estimates are constructed, we report three studies that show that there is only a weak relationship between the language used to describe an event and people's duration estimates of the event. Further, we address whether the language used directly influences the duration estimates. The findings show that the language used does not cause differences in duration estimates.

CONSTRUCTING DURATION ESTIMATES

Duration estimation depends on coding, storage, and retrieval processes of long-term memory. Most theories have focused on the complexities of the event and how these are perceived. One influential theory is Ornstein's (1969) storage size hypothesis. According to this theory, duration estimation is proportional to the amount of information stored in memory about the event. This explains, for example, why musical scores with more notes are estimated as being longer than equivalently long scores with fewer notes (Ornstein, 1996). If individuals remember an event differently they may produce different estimates of its duration.

How people remember an event is complex. Information feeds into memory from many sources, not just the initial perceptions, which themselves are filtered and distorted as they pass through the cognitive system. Several metaphors for memory have been used over the years (Draaisma, 2001; Roediger, 1980). One that we favour comes from Neisser's (1967) *Cognitive Psychology*, although it is one that he no longer endorses. When his seminal textbook came out, an influential philosophy of memory was what Neisser calls the 'reappearance hypothesis'. According to this hypothesis memories are encoded and then simply fade. The fading image degrades, but some structure remains essentially intact and what is left can be reactivated if enough associations are activated. Neisser, and before him James, Bartlett and others, were very dissatisfied with this notion. Neisser stressed that memories are reconstructed. His metaphor (then) was of a palaeontologist, trying to construct a dinosaur from a few 'stored fragments' (i.e. fossilized bones) coupled with theories of palaeontology.² Memories can be reconstructed from perhaps a few

²Neisser notes that Hebb used this metaphor for focal attention: 'Hebb's (1949, p. 47) comparison of the perceiver with a palaeontologist, who carefully extracts a few fragments of what might be bones from a mass of irrelevant rubble and 'reconstructs' the dinosaur that will eventually stand in the Museum of Natural History' (1967, p. 94). Hebb's own description, 'a drawing or a report of what is seen tachistoscopically is not unlike a palaeontologist's reconstruction of early man from a tooth and a rib', is less ambitious than Neisser's elegant reconstruction.

fragments of the filtered perceptions of our cognitive system, from script base knowledge of the event type, and from information about the event that was not part of the initial encoding.

With regards to the last of the means of constructing a memory, Loftus and colleagues (e.g., Loftus *et al.*, 1978; Loftus and Palmer, 1974; Wright *et al.*, 2001) have shown how information can be subtly introduced to people and can affect the reconstructive processes. Most applicable to the current studies, Loftus and Palmer (1974, Experiment 1) showed participants several film clips of automobile crashes and asked them 'About how fast were the cars going when they [contacted, hit, bumped, collided, smashed] each other?' with one of the five verbs shown in the brackets. The mean speed estimates increased from 'contacted' through to 'smashed' demonstrating how even something as innocuous as the choice of verb in a question can alter people's responses. The reason is that the different verbs imply different speeds.

There is an important aspect of Loftus and Palmer's (1974) paper that often gets overlooked. In their second experiment, they used only two verbs, 'hit' and 'smashed'. They found that people asked with the 'smashed' were more likely to report (non-existent) shattered glass at the scene of the crash. This aspect is covered in most textbooks. What is not covered is that even controlling for the speed estimate, there was still an effect of verb used on the presence or absence of shattered glass in the participants' memories. This is important because it shows that the verb choice is not simply changing the speed estimate which in turn the participant is using to construct a memory. The verb choice is directly affecting the memory.

Burt and colleagues (for example, 1992, 1993, 1999; Burt and Kemp, 1991; Burt and Popple, 1996) have applied Loftus' findings to duration estimates. Because the speed of an action is inversely related to the duration the action will last, an increase in implied speed should be coupled with a decrease in duration estimates. In Burt and Popple (1996) participants watched a confederate come into a lecture and proclaim 'repent spawn of saint, and you, you nest of vipers, you can't keep me here' (p. 55). Participants were then asked 'How long did it take the person to [run, pass, walk] through the lecture theatre?' (p. 56) with one of three different verbs. The actual duration was 25 seconds. The mean for the 'run' group was 48 seconds, the 'pass' group 62 seconds and the 'walk' group 80 seconds. As with the Loftus and Palmer (1974) study, the verbs implied different speeds, and these produced very different estimates of the duration.

In Burt (1999, Experiment 2) participants watched a bank robbery and were asked to write an account of the crime. After writing a description of the event, they estimated the duration of the event. Participants' descriptions were content analysed. The number of action words and the total number of words for the description were calculated. The results showed that the use of many action words was associated with lower estimates of the event duration when tested immediately after viewing the event.³ If people used more 'action words' they tended to say that the bank robbery lasted for less time. Burt (1999) concluded that 'causation is from the construction of a narrative to describe an event to an estimate of the event's duration' (p. 353). Pictorially this is represented in Figure 1(a) the person's memory is used to create an event description and aspects of this description influence the estimated duration. If true this has very important implications because by altering the way

³There was a group that did the description and duration estimation after watching the video and a group who did the description and estimation the next day. The relationship between action words and duration estimates only held for the group tested soon after the event. Because of this, participants in our studies give their estimates shortly after encoding the event.

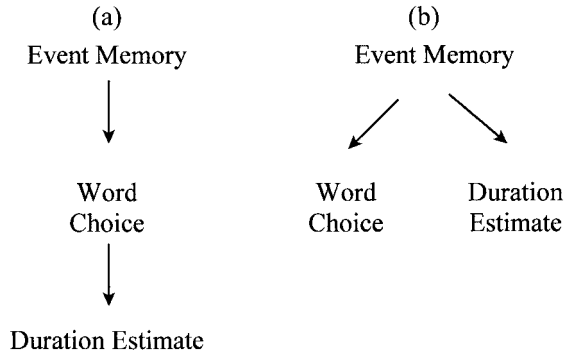


Figure 1. Two hypotheses about the relationship among memory for an event, language used to describe an event, and the duration estimate of the event. In (a) memory influences how the event is described which directly influences the duration estimate. In (b) the relationship between language used and duration estimation is spurious, both being influenced by memory.

people describe an event, you could alter the duration estimate. Let us consider the Oklahoma Bombing. If Johnston had spoken to a journalist before speaking to a police officer her duration estimate may have been shorter, and therefore more consistent with the timing and location of McVeigh's arrest.

Given the potential importance of Burt's (1999) conclusions, it is important to examine alternatives and look at his methods closely. We felt Burt's (1999) conclusions were not necessarily implied by his data, and also had a methodological concern. The problem is straightforward. It is difficult to reach causal conclusions of the type depicted in Figure 1(a) on the basis of correlational data (Holland, 1986). The association between words and duration estimates could be *spurious* (Simon, 1954): differences in the memory of event causing differences in both the description and the duration estimate (Figure 1(b)).⁴ In the following studies we use experimental procedures in an attempt to alter the descriptions of different events. If it is the descriptions, *per se*, that affect duration estimates, then our manipulation should also affect duration estimation, with descriptions using more action words leading to shorter duration estimates. If the relationship is spurious, as depicted in Figure 1(b), then we would expect no causal link between the description of events and the duration estimate.

Burt (1999, p. 350) suggests that it is 'individual differences in vocabulary [used to describe an event] may affect estimated event duration'. Recall that when describing an event a memory does not simply reappear as had been suggested by many of the early British empiricist philosophers (see Neisser, 1967, for discussion). It is reconstructed and people will differ in how their memories of the event are reconstructed. These differences may coincide with different beliefs about the duration of the event and may also lead to different words used to describe the event. It is possible that if a person remembers a lot about an event, and therefore according to Ornstein (1969) will tend to overestimate the duration, that they will also give more detailed event descriptions. Similarly, if a person

⁴Burt (1999) acknowledges that it could be that the event's estimated duration is influencing the word choice in the event descriptions. While examples could be conjured up that show this is possible in some specific situations, we agree with Burt that as a rule it is 'very unlikely' (p. 353).

recalls fast actions, which according to Burt and Popple (1996) will lead to underestimation, then they may also include words that imply faster action in their descriptions. This model is implied in Figure 1(b).

There is a third possibility. It is clear from Burt's, and other people's, studies that duration estimates are related to the way in which questions are asked (Burt and Popple, 1996), what the people say in the scenes (Burt and Popple, 1996), how long events typically last (Burt, 1993), the duration of the event (Yarmey, 2000), whether the event-type always has the same duration (Yarmey, 2000), and a host of other aspects. However, it may be that the interplay among language use and duration estimation is not as strong as is thought. Burt (1992) found that the knowledge someone had about an event did not predict duration estimates. Further, some amnesiac cases show a dissociation between memory and temporal judgements more generally (for example, Sirigu and Grafman, 1996). Perhaps, the way people describe events is not related to duration estimates. The correlation is neither direct (Figure 1(a)) nor spurious (Figure 1(b)), but is non-existent (or at least very small). Burt's (1999) conclusions were based on finding a correlation in only one group and an r of -0.41 ($p=0.05$) which equates with an r^2 of 17%. Using Steigler and Fouladi's (1992) R2 program, the 95% confidence for r^2 goes from 0% to 48%. Clearly further investigations are warranted.

Our methodological concern was that in calculating the number of action words, Burt (1999, Experiment 2) did not differentiate among them. As Burt points out, there was great variability in the choice of action words in his study. For some participants in Burt's study, the robber 'burst' into the bank, while for others he simply 'entered'. As it was the difference in implied speed that was the manipulation in Burt and Popple (1996) and Loftus and Palmer (1974), we felt this should be addressed. In our studies the quantity of words used and their average amount of implied action are measured separately.

Besides our interest in the theory behind duration estimation, we are also concerned with applied issues. After viewing a crime, eyewitnesses might be asked to describe what they saw to a police officer, to a journalist, to friends, and to family. The way in which you describe the event is likely to differ depending on to whom you are describing it. We hypothesize that the way in which a newspaper journalist might describe a crime would differ from how a police officer would. In particular we would expect a journalist, particularly if they worked for a 'tabloid' newspaper, to use more sensational action words. In Experiments 1 and 2 participants view crime scenes, are instructed to write descriptions in different styles, and then estimate the duration of the events. This allows us to see if manipulating the description style alters the duration estimates.

In Experiment 3 we eliminate the memory components of Figures 1(a) and 1(b) by having participants read descriptions of events which they have not seen and trying to estimate the duration from these. This is of additional applied interest. There are many ways in which people learn about events other than experiencing them. For example, in court hearings, eyewitness reports and police statements are often used to describe events to jurors. In these situations a crime is usually described in a formal manner. The media also provide a major source of information about events for most people. In particular tabloid newspapers in the UK are the principal source of information for many people. A recent survey (NRS, 1997, <http://www.nrs.co.uk/topline/9607to9706/newspapers.htm>) of daily newspapers showed that tabloids are read by about four times more people in the UK than the broadsheet newspapers. In particular *The Sun* was read by over 20% of people. *The Sun* tends to sensationalize stories, and probably best represents the stereotype of UK tabloids.

EXPERIMENT 1

Method

In order to estimate how many participants were needed in each condition a power analysis was conducted. The estimated effect size, based on Burt (1999) and conventions, was η^2 of 10%. Using equation 8.2.22 in Cohen (1988), this yields an f of 0.33. To achieve a power of 0.80 requires $n = 30$ per condition. This is approximately the size of the first-year undergraduate psychology programme at the University of Bristol. Eighty-seven first-year psychology undergraduates ($n = 29$ per condition) participated as part of a laboratory class (67 females and 20 males). The median age was 18 years.

A between-subjects design was used. The independent variable was which of three writing styles (personal style, tabloid style or police officer style) the participant was told to use to describe the event. In the personal condition participants were requested to describe the event in their own personal writing style. In the tabloid condition participants were requested to describe the event as if they were writing an article for a tabloid newspaper, like *The Sun*. They were asked to use dramatic and expressive language in the aim of creating a sensational and attention-grabbing story. In the police officer condition participants were asked to describe the event as if they were police officers writing a statement report. Thus formal, precise, and informative language was encouraged. Participants were randomly assigned to one of the three conditions with the constraint that equal numbers of participants were in each condition.

About fifteen minutes into a first-year undergraduate psychology class, a male confederate burst into the lecture theatre, walked across the front of the lecturer's table, stopped, and shouted: 'I'm sick and tired of psychology, its boring, you're boring, it's all boring.' The lecturer tried to make the man sit and calm down. The confederate continued to exclaim how boring psychology was. He ran towards the lecturer's table, closed an open briefcase that was on the table, and ran out of the fire exit with it while the lecturer shouting 'hey, hey' and following the confederate out. The event was timed on a stopwatch by the experimenter who was discreetly present in the class (it was also recorded). Neither the experimenter nor the confederate was known to any of the participants. Timing began from the moment the confederate entered the lecture theatre and finished when the lecturer exited the room. The event lasted 19.8 seconds. Both the confederate and the lecturer had received formal acting training at university level.

Immediately after the event, participants were handed a sheet of paper that asked them to describe the event they had just seen using one of the three styles. Participants did this and the lecture continued for ten minutes. Then, participants were given another sheet of paper asking them several questions about the event. The critical question asked people to estimate the duration of the event, from when the confederate came into the room, to when the lecturer left.

We were interested in two main aspects of these data. First, do the different instructions produce writing styles that can be differentiated using quantitative measures? In particular, the scales we produced measured the quantity of verbs and the total number of words, and also the implied action of the verbs. The second aspect is whether the different instructions lead to different estimates of the event duration. If the descriptions affect duration, then according to Ornstein's (1969) storage size hypothesis we would expect that longer descriptions would be judged as having longer durations. Further, according to Burt (1999)

the use of verbs that convey that the actors rushed during the event should produce shorter duration estimates.

Results

Our first task was to content analyse the descriptions. We counted the total number of words, the number of verbs, and the number of adverbs. There were very few adverbs used (most participants used none) and therefore we concentrated on the number of words and verbs. The number of words and the number of verbs are related measures of the quantity of the descriptions. For the verbs we wanted to get some idea about the action implied by each. A list of all 131 verbs used by participants was presented to five people (none of whom participated in any of the other studies) who rated the amount of action implied by each of these. The scale ranged from 1 to 10 with 1 being 'implying no action' and 10 being 'implying incredibly fast action'. The mean action rating scores from the five people's judgements were calculated for each verb. We then calculated the average action verb score for each participant by finding the mean score of verbs used by that person. Thus, we have three event description variables: number of words, number of verbs, and average action verb score. The means and confidence intervals for these variables are shown in Table 1. The first two of these are related to the detail of the description and to Ornstein's (1969) hypothesis. Because of the instructions we predicted that those writing in the police style would use the most words and verbs. This was observed ($F(2, 84) = 2.57, p = 0.08, MSe = 296.59, \eta^2 = 0.06$, for the number of words; $F(2, 84) = 6.72, p = 0.002, MSe = 18.03, \eta^2 = 0.14$, for the number of verbs). The next variable, the average action verb score, was highest in the personal and tabloid conditions ($F(2, 84) = 8.90, p < 0.001, MSe = 0.60, \eta^2 = 0.18$), as predicted.

Having found that our instructions to use different writing styles did produce quantitatively different descriptions, we now explore the relationship between conditions and duration estimates. Recall that the actual duration of the event was 19.8 seconds. The mean estimates for the different conditions are shown in Table 1. There is a trend for overestimation, although the true value is included in the confidence intervals for each

Table 1. Means and 95% confidence intervals for the three event description variables, the duration estimates and the natural logarithm of these, broken down by the experimental conditions for Experiment 1

	Condition					
	Personal		Police		Tabloid	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Number of words	64.28	(58.90, 69.66)	69.14	(61.08, 77.19)	58.90	(52.99, 64.80)
Number of verbs	12.69	(11.27, 14.11)	16.10	(14.17, 18.03)	12.45	(11.00, 13.89)
Average action verb score	3.09	(2.82, 3.36)	2.26	(2.03, 2.49)	2.87	(2.50, 3.24)
Duration estimate ^a	29.72	(17.56, 35.50)	41.21	(15.33, 55.39)	29.28	(17.43, 35.96)
ln of duration estimate	3.23	(3.02, 3.44)	3.33	(3.00, 3.65)	3.14	(2.87, 3.42)

^aBecause the raw duration estimates are highly skewed, standard procedures for calculating confidence intervals are inappropriate. Here, an asymmetric percentile *t* bootstrap (with 599 replications) is used (see Wilcox, 1997, for description and some alternatives) using specially written macros for S-Plus.

condition. The duration estimates were highly skewed (skewness = 2.49, $se = 0.26$). Several transformations (and procedures) were explored and all lead to the same conclusion. Here we report analysis based on the natural logs of these data (overall skewness = 0.38, $se = 0.26$). Table 1 gives the means and their 95% confidence intervals. These differences are not statistically significant ($F(2, 84) < 1$).

Figure 1 depicts two ways in which a correlation between word choice and duration estimation might arise. Figure 1(a) states that language use leads to differential estimation (Burt, 1999). Figure 1(b) states that the correlation is spurious due, perhaps, to variation in people's memories of the event. Both of these predict correlations within the conditions on the event description variables and duration estimates. In particular, following Ornstein (1969) the quantity of words should be positively associated with duration estimates. Following Burt and Popple (1996) we would expect the average action verb score to be negatively associated with duration estimates. From Burt (1999, Experiment 2) we might predict a negative correlation between number of verbs and duration. Correlations were estimated between the three variables from the content analysis (and action rating pilot study) and the logarithm of the duration estimates. This was done both for the sample as a whole and broken down within conditions. None of the twelve tests approached statistical significance ($min p > 0.20$) and all were small.

Discussion

Burt (1999) concluded that it was the words used to describe an event which influenced the estimate of the event's duration (Figure 1(a)). He based this on a negative correlation (in one condition) between the number of action words used and the duration estimate. However, this correlation could be spurious (Simon, 1954). It is possible, for example, that variation in the perception and memory of the event produces different event descriptions and different duration estimates (Figure 1(b)). There are several ways to disentangle these explanations. The simplest, conceptually, is to use an experimental design. In this first study we used a between subjects design and had people describe an acted out event in one of three styles. The different instructions did produce descriptions that differed in the predicted directions based on a content analysis.

If word use influences duration estimates, then the duration estimates should have differed among the conditions. In particular, as the police descriptions were longest, it might be predicted from Ornstein's (1969) storage size hypothesis that the duration estimates would also be the longest. Further, as the tabloid condition had the highest average implied action score, it might be predicted that this condition would produce the shortest duration estimates. As it turned out, there were no statistically significant differences for the duration estimates. Further, the correlations between the writing style variables and the duration estimates were all small. This counters both the hypotheses depicted in Figure 1.

The main finding of this study is that while our manipulation affected the event descriptions, there was a non-significant effect on the duration estimates. When making conclusions based on non-significant results, the power of the comparisons is critical. While an *a priori* power analysis was conducted, because individual differences in duration estimation can be large, a within-subject design would have more power than the between subjects design (Keren, 1993). In Experiment 2 we use a within-subject design. There are some further differences between Experiment 1 and Experiment 2. It is likely that some of the variation in the estimation may have been due to some students

paying more attention than others during the lecture. In Experiment 2 we tested people in a controlled laboratory environment. This is at the expense of using video clips as opposed to live acting. Further, multiple video clips were used.

EXPERIMENT 2

Method

Thirty Bristol university undergraduates took part in this study (11 females and 19 males). Ages ranged from 20–24 years with a median of 20 years. Participants were volunteers and unpaid. None had participated in Experiment 1 or rated the implied action of the verbs used in Experiment 1.

A within-subject design was used. Ten video clips of crimes from different television (mostly late-night cable) programmes were shown to each participant. All were filled with a series of actions. The first was a practice clip to ensure the instructions were understood. Participants were instructed how to use each writing style. We varied whether the clips were short (20 seconds), intermediate (50 seconds), or long (90 seconds). Counterbalancing was done such that each video clip was in each of the writing style conditions ten times and that each participant had one short, one medium and one long clip within each condition.

A Belinea MGA monitor with a screen dimension of 1280×1024 mm was used to display the stimulus events. The clips were edited using Ulead Media Studio and were shown using the package Matrox Rainbow Runner. The clips' dimensions were 320×240 mm. A description booklet was prepared for each participant. It contained a description of the different writing styles with an example. The instructions were essentially the same as those used in Experiment 1.

Participants were tested in pairs and were seated approximately one metre away from the monitor. The instruction sheet for the first phase of the study was handed out and after queries were dealt with and participants reported that they were ready, testing began. First the practice clip was shown. All participants described this in the tabloid style. If there were no further questions, the remaining clips were shown. Immediately after each clip had finished, participants wrote a description of the crime in the style of writing that the experimenter had told them to use. Participants wrote three descriptions in each of the three writing styles and the ordering of the styles was counterbalanced.

When the descriptions for all the clips had been completed, the instruction sheet for the second phase of the study was given to the participants. This told the participants that they would be required to answer questions about each of the crime clips, and encouraged participants to refer to their descriptions to help answer the questions. This encouragement was given to increase the chance that the writing styles would influence their duration estimates. The form of the questions was similar to that used in Experiment 1. Again, the critical question asked about the duration of the event.

Results

People's duration estimates were averaged over the short, medium and long durations for each writing style. When this was done there was one extreme outlier, whose score was more than 3.5 standard deviations above the mean for each of the three conditions. This

Table 2. Means and 95% within subject confidence intervals for the three event description variables and the duration estimates, broken down by the experimental conditions for Experiment 2. The total number of words and verbs are for all three events; the averages are per event

	Condition					
	Personal		Police		Tabloid	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Total number of words	210.10	(186.7, 233.6)	164.10	(146.5, 181.7)	185.40	(176.1, 194.6)
Total number of verbs	39.84	(33.42, 46.23)	36.63	(30.51, 42.72)	45.09	(37.71, 52.5)
Average action verb score	3.27	(2.98, 3.57)	2.98	(2.57, 3.38)	2.99	(2.64, 3.34)
Average duration estimate	42.65	(35.26, 50.03)	48.16	(42.85, 53.48)	49.40	(42.37, 56.42)

participant was excluded from analysis. No other score was more than two standard deviations from the mean. The resulting distributions did not need to be transformed. In the remainder of this section we follow the same order as in the results section of the first experiment. First, we discuss the content analysis of the descriptions and compare them across the different conditions. Next, we test if there are differences among conditions with respect to the duration estimates. Finally, we look at relationships between the variables from the content analysis and the duration estimates.

As with Experiment 1, all the event descriptions were coded for the total number of words and total number of verbs (and also adverbs, but as with Experiment 1, these are not included because there were too few). Another group of five participants, none of whom took part in any other part of this research, rated the verbs on a 1 to 10 scale for implied action. Average action verb scores were calculated in the same way as in Experiment 1. Table 2 shows the means and 95% confidence intervals for these. As the comparisons are within-subject, the 95% confidence intervals are within-subject intervals (see Wright, 1997).

The pattern of results is different from that found in Experiment 1. The manipulation of writing style instructions did make a difference, but the most striking differences are between the personal style and the others. There was a statistically significant difference for the number of words ($F(2, 56) = 4.71$, $p = 0.01$, $MSe = 3266.22$, $\eta^2 = 0.14$). As the stimuli used here were very different from those used in the first experiment, finding a different pattern of results is not that surprising. The important point is that the instructions did produce differences which could lead to differences in duration estimates. The duration estimates are given on the last line in Table 2; the differences among groups are non-significant ($F < 1$).

Table 3 gives the correlations, and their associated pairwise p -values, between duration estimates and the event description variables. Given the number of tests, and taking into account the problems for both Type 1 and Type 2 errors with multiple comparisons, p -values between 0.01 and 0.10 should be viewed cautiously. None of the comparisons were significant at $\alpha = 0.01$. Some were at $\alpha = 0.10$ and in the predicted directions. The number of words was positively associated with duration for two of the conditions and average action verb score was negatively associated with duration for the police condition. We must stress, however, that the main finding of these results is that none of the effects appear consistent or strong.

Table 3. The correlations and pairwise p -values comparing duration estimates and the writing style variables. If adjustments were made to the p -values to account for the increased Type 1 error associated with multiple comparisons, none of these would be significant at $\alpha = 0.05$

		Writing style variables		
		Number of words	Number of verbs	Average action verb score
Personal	r	0.41	0.29	-0.14
	p	0.03	0.13	0.48
Police	r	-0.05	0.33	-0.39
	p	0.80	0.09	0.04
Tabloid	r	0.32	0.08	0.17
	p	0.09	0.70	0.39

Discussion

The main result of Experiment 2 is that altering description style does not produce significantly different duration estimates. This replicates the findings of Experiment 1, using a within-subject design and multiple events, but is in contrast to Burt's (1999) hypothesis. Also consistent with Experiment 1 is that the correlations between the event description variables and duration estimates were fairly low. Increasing sample size is likely to produce statistically significant results, but the magnitudes of these effects are likely to be small.

We chose the tabloid and police officer styles for both theoretical and applied reasons. The theoretical reason is that we hypothesized that they would produce different characteristics which map onto Ornstein's (1969) and Burt's (1999) hypotheses about how duration estimates are made. The applied reason is that these are important sources of information for the general public. With respect to tabloid newspapers, they are read by most people in the UK. Our interest with the police style was because jurors, and others involved in criminal cases, will often hear a police report of an event. In these cases the people hearing about the event are likely not to have experienced the event. The question we ask in our third study concerns people's duration estimates when they have not experienced the event.

EXPERIMENT 3

Method

Fifty eight participants took part in this study, 23 females and 35 males. Ages ranged from 18 to 54 years. Participants were paid £2.50 and were recruited on the Bristol University campus. None of the participants had taken part in the previous two experiments (or the two action rating studies).

A within-subject design was used and the independent variables were the same as in Experiment 2. Each person in this study was given one set of descriptions from a person who took part in Experiment 2. The descriptions were all transcribed word-for-word, although the spelling was corrected. By using these description sets we in effect have the same counterbalancing as used in Experiment 2. Each set of Experiment 2 descriptions was used for two participants in this study. All the description sets of Experiment 2, except

for the one outlier who was excluded, were used to avoid any potential sampling bias. We used 58 participants, as opposed to just 29, based on a power analysis. Burt (1999) found a negative correlation of 0.41. From the results of Experiments 1 and 2 it is difficult to argue against at least some of this correlation being spurious. We set an *a priori* effect size of half this shared variance which yields an effects size of 0.29. An *n* of 58 results in a power of approximately 75% (Cohen, 1988, Table 3.3.2).

Participants were tested individually and an instruction sheet describing the format of the study was given to them. This informed participants that the written descriptions of nine events were made by participants in a previous study and that they were to read each description carefully, paying particular attention to the pace of events described, and to try to imagine these events. They were asked some questions after reading each of the descriptions. As with the previous two studies the critical question asked about the duration of the event being described.

Results and discussion

Our main interest was whether the different writing styles produced different duration estimates. This can be looked at in three ways. First, do the three conditions produce different duration estimates? Second, do the event description variables predict duration estimates? Finally, it might be that there is something idiosyncratic in the descriptions, not covered by the variables we found in our content analysis, that is predictive of duration estimates. To assess this we test for correlations between the person who wrote the description (Experiment 2 participants) and the estimates from the participants in this study.

The duration estimates were varied and were skewed. The skewness was 4.37, 2.69 and 3.51 (*se* = 0.31 for each) for the personal, police and tabloid conditions respectively. The natural logarithm was taken, reducing skewness to 0.46, 0.24 and 0.23 (*se* = 0.31 for each). These transformed variables will be used in the analyses.

The means (and 95% within-subject confidence intervals in parentheses) for the three conditions are: 4.57 (4.26, 4.88) for personal style, 4.61 (4.36, 4.86) for police style and 4.91 (4.61, 5.21) for tabloid style. A repeated-measures ANOVA confirmed that these differences are statistically significant ($F(2, 114) = 4.12$, $p = 0.02$, $MSe = 56.03$, $\eta^2 = 0.07$). However, the effect is not large and there is much variation within the conditions.

We felt a more sensitive test of the influence of writing style on duration estimates would be seeing how the event description variables (i.e. the number of words, the number of verbs and the average action verb score) related to the duration estimates. As there are three conditions and three variables for each description, we ran nine bivariate correlations. Only one was statistically significant (average action verb score for the police condition, $r = 0.27$, $p = 0.04$) and there were no clear patterns among the correlations. Given that the one significant effect is in the opposite direction as predicted (the higher the implied action, the *shorter* the duration should be) and any adjustment to control for Type I error with multiple comparisons will make this effect non-significant, we conclude that these variables are not good predictors of duration estimates.

Finally, we felt that the way in which the participants from Experiment 2 wrote their descriptions might convey some aspects of the duration that were more subtle than we picked up with our content analysis. Therefore, we used the times from those participants to predict the estimates here. None approached significance. The most sensitive test of

words influencing duration estimates is if the choice of words to describe an event in Experiment 2, when given to two different people, produced correlated duration estimates. None of the correlations approached significance.

In summary, participants were given nine different descriptions of events. Some of these were written in another person's own style, some in the style of a police officer and some in the style of a tabloid reporter. The participants then had to estimate the duration of the event being described. While there were a couple of statistically significant differences based on characteristics of the descriptions, overall, the main conclusion is that the descriptions did not have a large influence on the duration estimates. While we could focus on the few significant results, given the lack of systematic pattern of findings, we wish to be cautious in our interpretations.

GENERAL CONCLUSION

Duration estimation is important, particularly in a forensic context as is evident from the Oklahoma bombing case. Germaine Johnston reported seeing two men, one of whom looked like McVeigh, outside the Federal Building. She estimated that this was 28 minutes after the bomb exploded. McVeigh was arrested 77 miles away 50 minutes after this sighting. This would have been difficult and was used to discredit Johnston's sighting. But her duration estimate could have been errant. There are individual differences in duration estimates. Understanding how people make estimations may enable predictions of which estimates are accurate and possibly create interviewer guidelines to increase accuracy. If it had actually been only a few minutes between the explosion and Johnston's sighting then McVeigh could have quite easily been found 77 miles away.

In a series of papers, Burt and colleagues (for example, 1999, Burt and Popple, 1996) have explored the relationship between the language used to describe an event and duration estimation. In particular, Burt (1999) argues that the language used to describe an event affects the duration estimation of the event. This was based on a significant correlation (in one condition) between the number of action words somebody used to describe an event and the person's duration estimate. Figure 1(a) gives a schematic representation of Burt's (1999) conclusion and this has important consequences. However, this association could have arisen for several reasons and one possible alternative is depicted in Figure 1(b). It shows that differences in people's reconstructive memory of the event might produce differences in how the event is described and differences in duration estimation. The association between the description and estimation is said to be spurious (Simon, 1954). A third possibility is that the relationship between the description and estimation is not as strong as suggested from Figure 1. This is what our data suggest.

Our first two studies were designed to compare the two models depicted in Figure 1. By using an experimental manipulation designed to alter the words used to describe the event we were able to discriminate the two models. According to Figure 1(a) the manipulation would alter the duration estimates. According to Figure 1(b) it would not alter the duration estimates. In both studies the differences among the groups were nonsignificant ($F_s < 1$), which goes against Figure 1(a). In general, more of the effects are in the correct direction of Burt's hypothesis, but certainly there is no consistent pattern. While a small effect may still exist, further research is necessary to validate Burt's claim.

Both Figure 1(a) and 1(b) predict that within conditions the way in which people describe an event will be associated with their duration estimates. From Ornstein's (1969)

storage size hypothesis, we might predict that the number of words used to describe an event would be positively correlated with the duration estimates. Further, from Burt and Popple's (1996) description of how verb choice affects duration estimates, we predicted the higher the average action verb score, the shorter the duration. In Experiments 1 and 2, 21 correlations were conducted and only two reached statistical significance at $\alpha = 0.05$. From this we must conclude that the relationship between the writing style characteristics that we measured and the duration estimates is weak at best. This counters Burt's (1999, Experiment 2) finding a correlation of $r = -0.41$. However, the number of action words, his measure of writing style, is different from our measures.

Figures 1(a) and 1(b) each give 'Memory' a dominant role. Perhaps the apparent lack of predictive power for the descriptions is because people's memories for the events are too dominant compared with language. In the third experiment we wanted to eliminate this possibility. We gave the descriptions from Experiment 2 participants to a new sample of participants who did not see the events. There was now a difference in conditions, the 'tabloid' descriptions had the longest duration estimates. However, the effect was not large. As there were many differences within each condition, we felt the more powerful comparisons would be for the relationships between the description characteristics and the duration estimates. The results were similar to those found in the first two experiments. There was only one significant correlation out of nine, and it was not in the predicted direction. Again, this does not bode well for language use having a large impact on duration estimation.

In three studies the relationship between the words people use to describe an event and people's duration estimates of the events was investigated. Overall, none of the effects we found were large. In particular, our experimental manipulation, having participants write in different styles, did not produce significantly different duration estimates. This suggests that language use does not have a direct effect on duration estimates, or at least that the effects are not large. Further, correlations between characteristics of the description and the duration estimates were very low and showed no obvious patterns. This research demonstrates that some relatively simple hypotheses about the relationship between event descriptions and duration estimation might not be tenable.

ACKNOWLEDGEMENTS

We thank Lisa Babalis, Melanie Hall, and John Rose for help on coding the descriptions, preparing the stimuli, and acting in our first study, respectively.

REFERENCES

- Burt CDB. 1992. Reconstruction of the duration of autobiographical events. *Memory & Cognition* **20**: 124–132.
- Burt CDB. 1993. The effect of actual event duration and event memory on the reconstruction of duration information. *Applied Cognitive Psychology* **7**: 63–73.
- Burt CDB. 1999. Categorisation of action speed and estimated event duration. *Memory* **7**: 345–355.
- Burt CDB, Kemp S. 1991. Retrospective duration estimation of public events. *Memory & Cognition* **19**: 252–262.
- Burt CDB, Kemp S. 1994. Construction of activity duration and time management potential. *Applied Cognitive Psychology* **8**: 155–168.

- Burt CDB, Popple JS. 1996. Effects of implied action speed on estimation of event duration. *Applied Cognitive Psychology* **10**: 53–63.
- Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd edn). Erlbaum: Hillsdale, NJ.
- Draaisma D. 2001. *Metaphors of Memory*. Cambridge University Press: Cambridge.
- Holland PW. 1986. Statistics and Causal Inference. *Journal of American Statistical Association* **81**: 945–960.
- Keren G. 1993. Between- or within-subject design: a methodological dilemma. In *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Keren G, Lewis C (eds). Erlbaum: Hillsdale, NJ; 257–272.
- Loftus EF, Miller DG, Burns HJ. 1978. Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory* **4**: 19–31.
- Loftus EF, Palmer JC. 1974. Reconstruction of automobile destruction: an example between language and memory. *Journal of Verbal Learning and Verbal Behaviour* **13**: 3–13.
- Loftus EF, Schooler JW, Boone SM, Kline D. 1987. Time went by so slowly: overestimation of event duration by males and females. *Applied Cognitive Psychology* **1**: 3–13.
- Memon A, Wright DB. 1999. The search for John Doe 2: eyewitness testimony and the Oklahoma bombing. *The Psychologist* **12**: 292–295.
- Neisser U. 1967. *Cognitive Psychology*. Appleton-Century-Crofts: New York.
- Ornstein RE. 1969. *On the Experience of Time*. Penguin Books: Harmondsworth.
- Ornstein RE. 1996. *The Psychology of Consciousness* (2nd edn). Penguin Books: New York.
- Roediger HL. 1980. Memory metaphors in cognitive psychology. *Memory & Cognition* **8**: 231–246.
- Schacter DL. 2001. *The Seven Sins of Memory: How the Mind Forgets and Remembers*. Houghton Mifflin Company: New York.
- Simon HA. 1954. Spurious correlation: a causal interpretation. *Journal of American Statistical Association* **49**: 467–479.
- Sirigu A, Grafman J. 1996. Selective impairments within episodic memories. *Cortex* **32**: 83–95.
- Steiger JH, Fouladi RT. 1992. R2: a Computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research Methods, Instruments, and Computers* **4**: 581–582.
- Wilcox RR. 1997. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press: San Diego, CA.
- Wright DB. 1997. *Understanding Statistics: An Introduction for the Social Sciences*. Sage Publications: London.
- Wright DB, Loftus EF, Hall M. 2001. Now you see it; now you don't: inhibiting recall and recognition of scenes. *Applied Cognitive Psychology* **15**: 471–482.
- Wright DB, Self G, Justice C. 2000. Memory conformity: exploring misinformation effects when presented by another person. *British Journal of Psychology* **91**: 89–202.
- Yarmey DA, Matthys E. 1990. Retrospective duration estimates of an abductor's speech. *Bulletin of the Psychonomic Society* **28**(3): 231–234.
- Yarmey AD. 2000. Retrospective duration estimations for variant and invariant events in field situations. *Applied Cognitive Psychology* **14**: 45–57.
- Zakay D, Block RA. 1997. Temporal cognition. *Current Directions in Psychological Science* **6**: 12–16.