Running head: FLASHBULB MEMORY CONSISTENCY OVER LONG DELAYS

Consistency of flashbulb memories of September 11 over long delays: Implications for consolidation and wrong time slice hypotheses

Lia Kvavilashvili [1], Jennifer Mirani [1], Simone Schlagman [1], Kerry Foley [2] & Diana E. Kornbrot [1]

[1] University of Hertfordshire, UK

[2] University of Leicester

Address for correspondence:

Lia Kvavilashvili

School of Psychology

University of Hertfordshire

College Lane

Hatfield, Herts, AL10 9AB

United Kingdom

Tel. +44 (0) 1707 285121

Fax +44 (0) 1707 285073

Email: L.Kvavilashvili@herts.ac.uk

Abstract

The consistency of flashbulb memories over long delays provides a test of theories of memory for highly emotional events. This study used September 11, 2001 as the target event, with test-retest delays of 2 and 3 years. The nature and consistency of flashbulb memories were examined as a function of delay between the target event and an initial test (1-2 days or 10-11 days), and the number of initial tests (1 or 2) in 124 adults from the general population. Despite a reliable drop in consistency over the long delay periods, mean consistency scores were fairly high and the number of memories classed as 'major distortions' was remarkably low in both 2003 (9%) and 2004 (7%). The results concerning memory fluctuations across the re-tests and the qualitative analysis of 'major distortions' are consistent with the *wrong time slice hypothesis* which explains the development of distortions by hearing the news from multiple sources on the day of the flashbulb event (Neisser & Harsch, 1992). However, no support was obtained for the *consolidation hypothesis* (Winningham et al., 2000): memories of participants who were initially tested 10-11 days after September 11 were not more consistent than memories of participants tested 1-2 days after the event. In addition, the number of initial tests in September 2001 (one or two) and self-reported rehearsal did not have any beneficial effects on consistency. Together, these findings indicate that flashbulb memories may be formed automatically and consolidated fairly soon after an emotional event.

Keywords: Flashbulb memories, September 11, emotional memories, consolidation hypothesis, wrong time slice hypothesis

Consistency of flashbulb memories of September 11 over long delays: Implications for consolidation and wrong time slice hypotheses

Some events produce vivid and detailed memories that can stay with us for many years (e.g., a first date or a car accident) whereas other memories are less detailed and easily forgotten as time goes by. What makes some events more memorable than others? What is the role of emotion and rehearsal in the formation and maintenance of these vivid memories? Moreover, if something is vividly remembered with considerable detail, does this necessarily mean that the memory is veridical or resistant to distortion?

One area of study that has addressed these fundamental questions over the past 30 years is 'flashbulb memories'. These have been defined as particularly vivid and long lasting (autobiographical) memories "for *circumstances* in which one first learned of a very surprising and consequential (or emotionally arousing) event" (p. 73, Brown & Kulik, 1977, our italics). It has been customary to study these memories via unexpected and dramatic public events as, for example, the assassination of President John F. Kennedy (Brown & Kulik, 1977), the explosion of the space shuttle Challenger (Neisser & Harsch, 1992) or the resignation of British Prime Minister Margaret Thatcher (Conway, Anderson, Larsen, Donnelly, McDaniel et al., 1994). The important feature of these studies is that they do not examine one's memories for the details of the original event itself but for the so-called reception event - one's personal circumstances in which the news was first heard.

Initial studies of flashbulb memories concentrated on hearing the news about the assassination of John F. Kennedy many years after the event (Brown & Kulik, 1977; Winograd & Killinger, 1983; Yarmey & Bull, 1978). For example, in a seminal paper, Brown and Kulik (1977) reported that people were able to recall at least one of the six so called "canonical categories" about this reception event (the location, activity one was engaged in, source of news or informant, own emotion, others emotion and immediate

aftermath). Many participants also recalled irrelevant details such as "the weather was cloudy and grey", or "we all had on our little blue uniforms". Brown and Kulik (1977) found these results extraordinary given that autobiographical memories of ordinary events are less specific and tend to be forgotten within few months (Brewer, 1988; Larsen, 1992).

According to Brown and Kulik (1977), flashbulb memories are encoded by a special brain mechanism that switches on automatically whenever the levels of surprise and importance or consequentiality exceed a certain “threshold’. Although the resulting memory trace is not an exact (photographic) copy of the reception event, it is nevertheless fairly detailed and virtually unsusceptible to any decay or reconstruction for many years. Brown and Kulik (1977) emphasised the evolutionary importance of this biological "Now Print" mechanism (originally postulated by Livingston, 1967) that may have been crucial for survival in circumstances when one had to remember the details of potentially life threatening events (e.g., the time and location of first appearance of the rival tribes).

Although some of the Brown and Kulik’s (1977) initial ideas have been challenged, flashbulb memory itself remains an important and expanding area of research (Conway, 1995; Luminet & Curci, 2009; Pezdek, 2003b; Winograd & Neisser, 1992). However, the progress in this area has been relatively slow due to methodological difficulties of studying this phenomenon, contradictory findings and the scarcity of public events that would have a similar impact on the majority of tested samples (Brewer, 1992; Kvavilashvili, Mirani, Schlagman, & Kornbrot, 2003; Wright & Gaskell, 1995). In this respect, the tragic events that unfolded in New York on September 11, 2001 provided researchers with a new and unique opportunity to study the nature and mechanisms of flashbulb memories. In terms of surprise, emotional shock and consequentiality for the international community this event seems to surpass any other previously studied public event (see Luminet, Curci, Marsh, Wessel, Constantin, et al., 2004; Pezdek, 2003a;

Shapiro, 2006; Walters & Goudsmit, 2005). There are already signs of renewed interest in this area in terms of several recent publications (e.g., A.R.A. Conway, Skitka, Hemmerich, & Kershaw, 2009; Curci & Luminet, 2006; Ferré Romeu, 2006; Hirst, Phelps, Buckner, Budson, Cuc, et al., 2009; Niedzienska, 2004; Schmidt, 2004; Shapiro, 2006; Talarico & Rubin, 2003; 2007; Weaver & Krug, 2004; Wolters & Goudsmit, 2005) as well as a special issue of Applied Cognitive Psychology dedicated to memories of September 11 (Pezdek, 2003b) (see also Luminet & Curci, 2009).

Unlike Brown and Kulik (1977), who took their participants' memory descriptions at face value, most of the subsequent research on flashbulb memories has concentrated on the issues of (a) consistency of flashbulb memories and (b) whether there is a special mechanism for encoding and retaining these memories. The consistency of flashbulb memories is usually assessed via a test-retest method which involves comparing participants' memory reports obtained soon after the event (preferably within the first 48 hours) and then again at some later date (e.g., after several months or even years). It is assumed that if test-retest scores are very high over long time delays then this should be indicative of special encoding mechanism.[1]

Unfortunately, research conducted on test-retest consistency scores since Brown and Kulik's (1977) original paper has resulted in contradictory findings. Two early and influential studies that showed this contrasting pattern of results were conduced by Neisser and Harsch (1992) and Conway et al. (1994). Neisser and Harsch (1992) interviewed 44 undergraduates about their flashbulb memories of the Challenger explosion one day after the event and again after almost three years. The comparison of memory scores revealed a high degree of inconsistency between participants' reports obtained immediately after the event and those at re-test (after 32-34 months). The analysis of inconsistent reports also provided initial support for the *wrong time slice hypothesis:* because most participants had

watched the TV coverage of news some time after the explosion, several participants incorrectly assumed at re-test that it was the TV they first heard the news from. Neisser and Harsch (1992; see also Neisser, 1982) concluded that flashbulb memories are not necessarily consistent, and are ordinary memories that have been preserved by frequent rehearsal rather than by the operation of some special encoding mechanism (for similar views see Cubelli & Della Sala, 2008; Curci, et al., 2001; McCloskey, 1992; McCloskey et al., 1988; Weaver, 1993; Wright, 1993; Talarico & Rubin, 2003; 2007; Winograd, 1992).

In contrast, a study conducted by Conway et al. (1994; see also Cohen, Conway & Maylor, 1994) on a large number of British participants using the same test-retest method produced findings that were more in line with Brown and Kulik's (1977) views. Indeed, memory scores for the resignation of Margaret Thatcher taken within the first two weeks and then 11 months after the event displayed remarkably high levels of consistency: Eighty-six percent of participants had consistent flashbulb memories despite the fact that very strict criteria were used to define a memory as flashbulb (see also Pillemer, 1994). Therefore, Conway et al. (1994) argued that flashbulb memories may "constitute a class of autobiographical memories distinguished by some form of preferential encoding" (p. 326).

Over the past decade evidence has started to accumulate in support of both positions. Contradictory findings have even emerged for the same public event like the terrorist attack on New York on 11 September, 2001. For example, in a study by A.R.A. Conway et al. (2009), the mean percentage of consistent responses to questions about source, activity, location and others present was quite high at 80% after a delay of 11 months (see also Curci & Luminet, 2006, Shapiro, 2006; Tekcan, Ece, Gülgöz & Er, 2003). In contrast, in a study of Hirst et al. (2009), the mean consistency score after a similar delay was only .63 (scores ranged from 0 to 1) (see also Lee & Brown, 2003, Smith, Bibi, & Sheard, 2003; Talarico & Rubin, 2003; 2007).

One possible reason for inconsistent findings is the absence of a standard methodology for collecting flashbulb reports and different coding schemes used by researchers (for further discussion, see Shapiro 2006). For example, following Neisser and Harsch (1992), several studies have assessed the canonical categories of time, location, activity, source and others present, whereas others have used different questions including one's own or the informant's emotion, clothing, aftermath, and so on. There is, however, growing evidence to show that responses to the latter questions are less consistent than to the key canonical questions about the location, activity and source (e.g., Christianson & Engelberg, 1999; Ferré Romeu, 2006; Hirst et al., 2009; Schmidt, 2004; Weaver & Krug, 2004). Therefore, the inclusion of these questions into the calculation of consistency scores will inevitably reduce the overall consistency scores (see e.g., Hirst et al., 2009 whose participants showed particularly low consistency for their own emotion). Similarly, some studies have used Neisser and Harsch's (1992) graded 3-point scoring system to code the consistency of flashbulb memories (see below), whereas others have used a variety of other systems, for example, a more simple 2-point coding scheme (where 0 is an inconsistent and 1 - a consistent response). These differences make it very difficult to compare findings across studies. The situation is further complicated by the lack of agreement on the level of consistency that is necessary for classifying a memory as a flashbulb. Since none of the flashbulb memory studies have reported 100% consistency in 100% of participants, the results of the study with the consistency levels of 85% can be reported either in support or against the special status of flashbulb memories, depending on the researchers' theoretical preferences.

Apart from methodological/conceptual issues, the studies also differ on a variety of other variables which may also affect the outcomes of a particular study and further contribute to the inconsistencies documented in the literature (e.g., the nature of flashbulb

events, participant samples, media coverage, the timing of tests and re-tests, and so on). All this poses a considerable challenge to current flashbulb memory research and emphasises the importance of studies that seek to explore the variables that may be involved in producing these inconsistencies. Take, for example, Winningham, Hyman and Dinnel's (2000) claim that one of the most critical factors in producing inconsistencies across the studies is the length of delay between the original event and initial documentation of the reception event (see also Neisser et al., 1996; Rubin, 1992). According to their *consolidation hypothesis,* most of the forgetting occurs in the first few days of the original event. After this the memory traces consolidate into a relatively permanent narrative account (see also Weaver & Krug, 2004). Therefore, participants interviewed within the first few days after the event should exhibit poorer consistency across test-retest sessions than those who were initially tested several weeks after the original event. Winningham et al. (2000) provided preliminary evidence in support of this hypothesis by showing that 8 weeks after O. J. Simpson's acquittal, participants had more consistent memory scores if they were initially tested one week after the announcement of acquittal than only 5 hours after this announcement (see also Schmidt, 2004).

In contrast, Schmolck, Buffalo and Squire (2000) suggested that it is the length of the delay interval between test and re-test, rather than the delay between the event and the initial test, that is a crucial factor in determining the outcome of any particular study. They argued that the consistency of flashbulb memories was quite good in several studies that used relatively short time delays of 6 to 12 months, whereas substantial forgetting and distortions were observed by Neisser and Harsch (1992) with a delay of 32-34 months. In order to test the hypothesis that significant qualitative changes in flashbulb memories may be occurring with longer delays, Schmolck et al. (2000) re-tested two groups of participants after 15 and 32 months from the announcement of the verdict of O.J.

Simpson's murder trial. While consistency was quite good after 15 months, major distortions were observed in the group with a 32-month delay. However, as pointed out by Horn (2001), one possible confound in this study was the announcement of a second verdict from O. J. Simpson's civil trial 16 months after the initial verdict. It is possible that participants in the 32-month delay condition were confusing their memories of these two separate events, hence the high levels of distortions observed in the study. Therefore, according to Horn (2001), "the question of whether flashbulb memories decay over time is still open" (p.180) (*cf*. Hirst et al., 2009).

The present investigation had three principal aims. First, we wanted to study the consistency of flashbulb memories after a long delay of almost 2 years for a highly consequential and emotive event - the terrorist attack in New York on September 11, 2001. If, as suggested by Schmolck et al. (2000), major qualitative changes occur in flashbulb memories after the first 12-15 months from the event, then high levels of inconsistency and distortions should be observed after 23-24 months from the reception event (in July/August 2003). Second, in order to test the consolidation hypothesis of Winningham et al. (2000), half of the participants in this study were initially tested on 12th and 13th of September (a short delay between the event and an initial test), and half on 21st and 22nd of September (a longer delay between the event and an initial test). If several days are necessary for an initial memory trace to consolidate into a stable narrative account, as stipulated by the consolidation hypothesis, then test-retest consistency scores of participants who were initially tested on 21-22 September should be reliably higher than participants who were tested on 12-13 September.

The consolidation of memory traces in the first few weeks after the reception event was further examined by having half of the participants in each delay condition tested again two weeks after their initial test in September 2001. If this re-test acted as rehearsal,

reactivating and further consolidating the newly formed memories, then participants who were tested twice shortly after September 11 would have better consistency scores at long delays than participants who were tested only once in September 2001 (see e.g., Coluccia, Bianco & Brandimonte, 2006). Having the additional re-test in half of the sample soon after the first test was also important for calculating the initial consistency scores for a very short 2-week delay from the first test, and allowed us to correctly assess the amount of forgetting that may have occurred between this initial re-test (when the memories were fresh) and subsequent re-test after 23-24 months. To our knowledge, only one previous study has obtained such initial consistency measures and compared them to consistency scores after delays of 1 month, 3 months and 1 year (see Weaver & Krug, 2004).[2] The results showed that a percentage of consistent responses was near ceiling after one week from the first test (96%) and dropped reliably to 81% at one-year re-test. It is however, unclear, what would be the rate of forgetting (i.e., drop in consistency) with a longer delay of two years that is not confounded by additional re-tests at 1- and 3-months.

The third major objective was to examine possible fluctuations in memory descriptions over long time delays and to assess the wrong time slice hypothesis of Neisser and Harsch (1992). To this aim, all participants were re-tested again in July/August 2004, almost three years after the reception event on September 11, 2001. Not only did this additional re-test allow us to examine a possible drop in consistency scores from summer 2003 to summer 2004, but it also gave us a unique opportunity to observe a fate of memories coded as 'major distortions' in summer 2003. Would participants stick to their distorted memory accounts in summer 2004 or would they revert back to their original accounts in 2001? The only three studies that have addressed this important issue have resulted in mixed findings. Thus, participants in Neisser and Harsch (1992), who had distorted memories after a delay of 32-34 months, produced the same distorted memories

again after a delay of 38-39 months. However, two recent studies of September 11 produced opposite results by showing some fluctuations in memories across re-tests (Hirst et al., 2009; A.R.A. Conway et al., 2009). The results of Hirst et al. (2009) are particularly interesting because they showed that only 40% of initially inconsistent memories remained inconsistent at second re-test after 35 months (in contrast to 82% of consistent memories that remained consistent). Approximately 28% of inconsistent memories reverted back to original reports and 32% memories, albeit inconsistent at second re-test, provided a different story from that of the first re-test. These interesting fluctuations of inconsistent memories across re-tests appear to provide some support for the wrong time slice hypothesis which stipulates that major distortions tend to occur because people hear the important news from several different sources throughout the day and, at re-test, incorrectly remember some other (but real) occasion of hearing the news instead of the first occasion. The results of A.R.A. Conway et al. (2009) and especially Hirst et al. (2009) appear to indicate that the first time memories are not necessarily (and permanently) replaced by memories of hearing the news on later occasion(s). In order to address this issue, we assessed the fluctuations of memory descriptions across the 2003 and 2004 re-tests and examined the content of memories coded as 'major distortions' in 2003 and 2004.

General methodological considerations

At each data collection point, a Flashbulb Memory Questionnaire, modelled after Conway et al. (1994) was administered to participants by telephone interview (see Christiansen, 1989, Davidson & Glisky, 2002, Davidson, Cook & Glisky, 2006, for a similar procedure). Participants had to first provide a brief, but detailed, memory description of their personal circumstances in which they first heard of the terrorist attack in New York. This was followed by participants answering five questions about the canonical categories of *time*, *location*, *activity*, *others present* and *source*. Finally,

participants had to provide ratings on several scales assessing such background variables as surprise, emotion, importance (personal and national), rehearsal, vividness, etc.

Test-retest consistency scores were calculated by comparing memory descriptions and answers to the five questions at initial test to those at subsequent re-test (see Conway et al., 1994; Neisser & Harsch, 1992). Memory descriptions and the answers to canonical questions were coded separately because they were deemed to rely on distinct retrieval processes: free recall and probed (or cued) recall, respectively. For coding the consistency of probed recall we used the coding scheme originally developed by Neisser and Harsch (1992) and their Weighted Attribute Score (WAS). This coding method has been used in a large number of studies (e.g., Conway et al., 1994; Cohen et al., 1994; Curci & Luminet, 2006; Davidson & Glisky, 2002; Hornstein, et al., 2003; Shapiro, 2006; Schmolck et al., 2000; Smith et al., 2003; Tekcan et al., 2003), and it allowed us to compare our results with previous findings of Schmolck et al. (2000) and Neisser and Harsch (1992), who obtained very low WAS after long delays of 32 and 32-34 months, respectively.

Unlike many flashbulb memory studies that use undergraduate students, participants were recruited from the general population. Although there were roughly equal numbers of young (aged 20-56 years) and old participants (aged 61-82 years), the data are presented on the entire sample as participants age did not correlate with any of the dependent variables reported in the paper. This decision was further justified by a study of A.R.A. Conway et al. (2009) on a large national random sample (N=687) which had approximately equal numbers of participants in four age groups (18-29, 30-44, 45-59 and 60-87) and did not find any correlation between participants' age and consistency scores of flashbulb memories of September 11 (see also Davidson et al. 2006, Davidson & Glisky, 2002, and Otani et al., 2000 for similar non significant results).

Method

Design

The design was a mixed factorial with two between subjects and one within subjects independent variables. The first between subjects factor was the delay between the reception event and the initial test (short vs. long). Half of the participants were tested 1-2 days after September 11 (short interval), and half were tested after 10-11 days (longer interval). The second factor was the number of tests in 2001 (one vs. two). Half of the participants were tested only once and half were tested again after two weeks from their initial test. All participants were contacted again for the final re-tests in summer 2003 and summer 2004, two and three years after the initial testing in September 2001. Therefore, the within subjects factor was the final retest delay (two years vs. three years).

Participants

A total of 168 British participants were initially tested in September 2001. They were recruited from an existing pool of volunteers from local community maintained by the first author and by contacting colleagues, relatives and friends of four researchers (first author and three research students).[3] Of these, 135 (80%) were re-tested in summer 2003. All participants were screened for cognitive functioning at the time of their 2003 interviews (for details see Kvavilashvili, et al., 2009). The data of four old participants with possible cognitive decline were excluded resulting in a sample of 131 participants. Of these, 124 (66 females, 58 males) were re-tested again in summer 2004. The mean age of the final sample was 53.12 (*SD*=20.55, range 20-81), and the mean number of years in education – 15.35 years (*SD*=4.48, range 8-28). Mean age and years in education did not differ as a function of independent variables as shown by the non-significant results of the 2 (delay of initial test: short, long) x 2 (number of tests in 2001: one, two) between subjects ANOVAs (both $F_s < 1$). For all participants English was their first language.

Materials

The Flashbulb Memory Questionnaire was divided into three sections (*cf*. Conway et al., 1994, Neisser & Harsch, 1992): (1) Participants had to provide a short but detailed narrative description about their personal circumstances upon hearing the news (i.e., free recall of the reception event); (2) Then they had to answer five canonical questions about the time (*when did you hear about the news*), the place (*where were you at the time*), the activity (*what were you doing*), the source of the news (*how did you find out*), and others present (*if not alone then indicate who else was present*) (i.e., probed recall of the reception event); (3) Finally, they had to provide ratings of various encoding and rehearsal variables on 10-point rating scales. Specifically, participants had to rate their levels of surprise, intensity of initial emotion, and intensity of stress later on in that day (1= not surprised/emotional, etc, 10= extremely surprised/emotional, etc.). They were also asked to rate how often they had been thinking about the terrorist attack (1=not at all, 10=all the time), and had to rate the vividness of their memory for the reception event (1=no image at all, 10= extremely vivid image, almost like normal vision). An identical questionnaire was re-administered to half of the sample two weeks from their initial test in September 2001.

The questionnaire that was administered to all participants in summer 2003 was also identical to the first questionnaire except that several new items were added. For example, participants had to provide confidence ratings for their memory description and for their responses to each of the five probe questions on a 10-point rating scale (1=merely guessing, not confident; 10=extremely confident). The section about various encoding and rehearsal variables contained two additional questions assessing perceived levels of personal and national importance of September 11 for participants when they first heard the news.[4] The question asking how much they had been thinking about the terrorist attack was changed to reflect the delay of two years ("How often have you been thinking/or being reminded of the terrorist attack in New York during the past two years?"). An

additional question assessed how frequently participants had rehearsed their memories of the reception context ("How often have you been remembering and/or thinking of your personal circumstances in which you heard of the terrorist attack in the past two years?").

The questionnaire that was administered to participants in summer 2004 was identical to the one administered in summer 2003, except for the two rehearsal questions: participants were asked to rate how frequently they had been remembering/thinking of September 11 and their personal circumstances in the past year instead of the past 2 years.

Procedure

Participants were individually contacted by one of four researchers by telephone on 12th and 13th of September or on 21st and 22nd of September, 2001. They were invited to take part in a study examining people's memories of how they first heard the news of a major public event such as the terrorist attack in New York. It was explained that participation was voluntary and that a few more interviews could follow in subsequent years. After obtaining oral consent from the participant, the Flashbulb Memory Questionnaire was administered over the telephone. Participants were asked to talk slowly and clearly into the phone so that the researcher could accurately record their responses. All participants complied with this request. On those few occasions when they did not, the researcher stopped them immediately, and repeated the request. This ensured that responses were recorded verbatim. Interviews lasted between 10 and 20 minutes.

Half of the participants were re-tested after 2 weeks from this initial interview. They were specifically asked to recall the reception event as they remembered it on that day rather than trying to remember the answers they gave in the previous interview. All participants were subsequently re-tested, after a delay of 23-24 months, in July/August of 2003, and after a delay of 35-36 months in July/August 2004. At the end of the interview

in 2003, participants completed three tests measuring their cognitive functioning and provided information about years of education.

Coding for consistency of probed recall

We used the coding scheme originally introduced by Neisser and Harsch (1992) and their Weighted Attribute Score (WAS). Participants' answers to each of the five questions (about time, location, activity, others present and source) at the re-test were assigned a score of '0', '1', or '2' depending on how *consistent* they were with the answers at the initial test. A score of '0' was assigned if participants said they could not remember or if they recalled information (e.g., 'my father') that was completely different from what they said at the initial test (e.g., 'my friend' in case of the source question). A score of '1' was assigned if participants provided either less specific information ('my friend' instead of 'my friend Jon') or slightly incorrect information (e.g., 'my friend Sam' instead of 'my friend Jon'). Finally, a score of '2' was assigned if participants provided either the same information at both tests (e.g., 'my friend') or the same information plus additional detail at the re-test (initially 'my friend' and then 'my friend Jon') (see Appendix 1 for details).

The total consistency score, derived from this coding scheme varies from 0 to 10. However, according to Neisser and Harsch (1992), correctly remembering *location*, *activity* and *source* has more weight than remembering *time* and *others present*, the less important attributes of flashbulb memories (see Tekcan et al., 2003 for providing direct empirical support for this idea and Shapiro, 2006 for further discussion). The WAS reflects this by assigning a maximum score of '2' for *location*, *activity* and *source*, and giving one bonus point if a participant's cumulative score for *time* and *others present* is '3' or more (out of a total possible 4). The resultant WAS can therefore vary from 0 to 7 with higher scores reflecting better test-retest consistency. Although identical results were obtained for total consistency and WAS, only the latter will be reported throughout this paper.

Coding for consistency of free recall

Participants' memory descriptions at re-test in 2003 were compared to those at initial test in 2001 and were classed into six possible categories: can't remember, major distortion, minor distortion, less specific, more specific and the same. This coding scheme was adopted because Neisser and Harsch's (1992) 3-point scheme does not distinguish *major distortion* from *can't remember* (both are coded as '0') or *minor distortion* from *less specific* response (both are coded as '1'). Thus, if participants could not remember, their response was categorized as *can't remember*. A memory description was classed as a *major distortion* if it was somewhat different (two or more attributes inconsistent, for example, activity and source) or completely different (all mentioned attributes inconsistent) from the original description. A memory description was deemed to contain a *minor distortion* if one of the canonical categories in the description was slightly incorrect (e.g., initially in my office at work and then in the staff room at work). Memory was coded as *less specific* or *more specific* if it contained less specific or more specific information about one or more canonical categories mentioned in the original description. If a memory contained the same canonical categories with the same level of specificity as in the original description it was classed as the *same* even if participants used different wording from the original. When coding memory descriptions, participants' answers to the specific questions were used to resolve any ambiguity and vice versa, i.e., all available information was utilised to obtain the most complete measures of memory consistency.[5] All the coding was carried out by several pairs of independent coders. The percentage of agreement varied, on average, from 85% to 100%, and the discrepancies were solved by discussion.

Results

The results will be presented in several sections reflecting the dependent variable analysed. Initially, we analysed a set of background variables to see if there were any

effects of independent variables (the time of initial testing and the number of initial tests in September 2001) on how the events were assessed by participants in terms of surprise, emotion, rehearsal, etc. We then examined participants' consistency scores in 2003 and 2004 separately for probed recall (participants' answers to the five questions) and free recall (memory descriptions). Additionally, for probed recall, we assessed a drop in the consistency scores over the 2- and 3-year delay periods in half of the sample who were initially re-tested after two weeks from their first test in September 2001. For free recall, we also examined the fluctuations of participants' memory descriptions across the two retest sessions and the content of major distortions. Finally, we calculated correlations between the background variables and the probed recall consistency scores. Unless otherwise specified the rejection level for all analyses was set at .05 and the magnitude of effects was measured by partial eta-squared (η^2). Furthermore, in all analyses of variance with repeated measures, if the sphericity assumption was violated, the reported *p* values were adjusted accordingly using Greenhouse-Geisser correction.

Background variables

All background variables were measured on 10-point rating scales (1=not at all, 10=extremely). The mean ratings of variables that refer to participants' initial reactions to the terrorist attack in September 2001 and variables that were collected in 2003 and in 2004 were entered into several 2 delay of initial testing (short vs. long) x 2 number of tests in 2001 (one vs. two) between subjects ANOVAs. No main effects or interactions were significant, therefore, the data on background variables are presented on the entire sample as a function of year of testing (2001 vs. 2003 vs. 2004). Means are presented in Table 1 together with the results of one way within subjects ANOVAs and effect sizes.

The results showed that the ratings of surprise and the vividness of memory image (for the reception event) were very high and remained stable over the three years as did the

ratings of stress, national and personal importance and confidence in the accuracy of free recall (memory descriptions) and probed recall (answers to five questions). However, some ratings changed reliably over time. For example, ratings of initial emotion showed a large increase from 2001 to 2003 ($p<.00001$) and then a small but reliable decrease from 2003 to 2004 ($p<.04$). However, the mean rating in 2004 was still reliably higher than in 2001 ($p<.00001$). Rehearsal of the September 11 event itself strongly decreased at each time point (all $p_s<.00001$). Conversely, rehearsal of personal circumstances of hearing the news (not assessed in 2001) reliably increased from 2003 to 2004 ($p<.0001$).

Consistency of probed recall (responses to the five questions)

In order to assess the consistency of probed recall, the mean Weighted Attribute Scores (range 0-7) were calculated by comparing participants' responses to five questions at their initial test in September 2001 with their subsequent responses in 2003 and 2004, respectively. The resultant consistency scores for 2003 and 2004 retests as a function of delay of initial testing and a number of tests in 2001 are presented in Figure 1 (see lines depicting data for 2003 and 2004 retests). These means were entered into a 2 delay of initial testing (short vs. long) x 2 number of tests in 2001 (one vs. two) x 2 year of re-test (2003 vs. 2004) mixed subject ANOVA with the repeated measures on the last factor. The only reliable effect was obtained for the year of re-test ($F(1,120)=6.34$, $MSE=.76$, $p=.01$, $\eta^2=.05$) with slightly better consistency scores in 2003 ($M=5.15$; $SD=1.58$) than in 2004 ($M=4.88$; $SD=1.72$). There were no reliable effects of delay of initial testing ($F<1$) and the number of tests in 2001 ($F<1$) as would be predicted by the consolidation hypothesis. All 2- and 3-way interactions were also non significant (all Fs<1).

For the 65 participants who completed two tests in 2001 we calculated additional consistency scores by comparing their responses to the five questions at the initial test in September 2001 to their responses obtained two weeks after initial testing. These 'initial'

consistency scores (obtained in 2001) were then contrasted with the 'subsequent' consistency scores obtained in 2003 and 2004 (these were the same as in previous analysis). Thus, the mean WAS in 2001, 2003 and 2004 were entered into a 2 delay of initial testing (short vs. long) x 3 time of re-test (2001 vs. 2003 vs. 2004) mixed ANOVA with the repeated measures on the last factor (see Figure 1, lines depicting data for 2001, 2003 and 2004 retests for participants who were tested twice in 2001). This analysis revealed a highly significant main effect of time of re-test, $F(2,126)=39.42$, $MSE=.95$, $p<.0001$, $\eta^2=.38$. Post hoc comparisons showed that the mean consistency scores in 2001 ($M=6.32$, $SD=.92$) were reliably higher than the consistency scores in both 2003 ($M=5.09$. $SD=1.38$) and in 2004 ($M=4.92$, $SD=1.58$) (both $p_s<.00001$). However, with only 65 participants rather than the entire sample, the difference between the 2003 and 2004 scores was not significant ($p=.25$). No other effects or interactions were significant (all $Fs<1$).

Although there was a substantial drop in the consistency scores from 2001 to 2003 and 2004, the mean WAS in 2003 and 2004 (in 65 participants and in the entire sample) are markedly higher than those in Neisser and Harsch (1992) for the Challenger explosion and Schmolck et al. (2000) for the O. J. Simpson's acquittal ($M= 2.95$ and $M=3.30$, respectively). The distribution of scores (0 to 7) in our sample in both 2003 and 2004 and in the study of Neisser and Harsch (1992) is also markedly different (see Figure 2).

Consistency of free recall (memory descriptions)

Participants' memory descriptions in 2003 and 2004 were compared to their initial memory descriptions in September 2001 and classed into one of the six response categories: *can't remember*, *major distortion*, *minor distortion*, *less specific*, *more specific*, and *same* as described above (see method). Figure 3 shows the percentage of participants (pooled across factors of delay of initial test and number of tests in 2001) whose memory descriptions in 2003 and 2004 fell into these six categories.[6] There were very few

participants who said they could not remember their personal circumstances (2% in 2003 and in 2004). The majority of memories (up to 60%) were without distortion but changed in specificity (became either more or less specific). More specific responses often arose because at re-tests participants tended to describe a larger time window (e.g., preceding events or aftermath) in addition to an exact moment when they heard the news.[7] There were relatively few participants who provided exactly the same number of details with the same level of specificity (7% and 6% in 2003 and 2004, respectively). About a quarter of descriptions contained a minor distortion with only one of the canonical categories being different/incorrect at the re-tests. Finally, the number of major distortions where participants had two or more inconsistent categories or provided a completely different story was also relatively small (9% in 2003 and 7% in 2004). The overall difference between the two re-tests in 2003 and 2004, as shown on Figure 3, was not significant, *Kendall's W*=.027, p=.096. However, post hoc comparisons showed that the percentage of *more specific* responses went down from 2003 to 2204, whereas the percentage of *less specific* responses went up. This effect was significant, *McNemar* χ^2= 6.25, p=.02.

Memory fluctuations and evidence for wrong time slice hypothesis

The percentages presented in Figure 3 do not provide any information about what happened to each individual memory description across the two retests in 2003 and 2004. Did they remain stable or did they fluctuate in terms of categories that they fell into? To answer this question, we constructed a 6 x 6 frequency table (Table 2), which shows the percentages of memories in each of the 6 response categories in 2004 as a function of response categories in 2003. This table shows a fair amount of fluctuation in each of the six response categories. Cohen's Kappa for 'agreement' in classifications of memory descriptions from 2003 to 2004 is .36 (with 95% confidence limits of .19 to .50), which is less than a 'moderate' level of agreement of .40 to .60 (see Landis & Koch, 1977). Take,

for example, 29 memory descriptions classed as a *minor distortion* in 2003. Only 17 (59%) remained in the same category in 2004, whereas 8 memories (27%) and 2 memories (7%) were coded as *less specific* and *more specific*, respectively. Importantly, only 2 memories (7%) were classed as *major distortion* in 2004. Similarly, out of 11 participants whose memories in 2003 were coded as *major distortion*, 7 participants (64%) remained in the same category in 2004 (six remembered the same distorted memory and one remembered a different distorted memory). However, four memories (36%) were re-coded in 2004 as *minor distortion* (18%), one memory became *same* (9%) and another *less specific* (9%).

We also used a more coarse grained classification to make our data more comparable to Hirst et al. (2009). Thus, all memories coded as *less specific, more specific* and *same* were pooled into a general category of broadly consistent memories, and *minor distortions*, *major distortions* and *can't remember* responses into a category of inconsistent memories. Table 2 shows that out of 82 consistent memories in 2003, 67 (82%) remained consistent and 15 (18%) became inconsistent in 2004. Out of 42 inconsistent memories in 2003, 29 (69%) remained inconsistent and 13 (31%) became consistent in 2004. There is some indication that the likelihood of memory becoming consistent in 2004 after being inconsistent in 2003, $p_1 = .31$, was somewhat higher than the likelihood of memory becoming inconsistent in 2004 after being consistent in 2003, $p_2 = .18$. The difference, $p_1 - p_2 = .13$, approaches statistical significance ($z = 1.52$, $p = .064$).

The qualitative examination of *major distortions* in 2003 provides important information about why participants' memory descriptions became distorted. They show that after hearing the news for the first time people heard the news again from multiple sources and at re-test they incorrectly remembered this subsequent (but real) occasion of hearing the news as the very first occasion they heard it. Relevant examples of memory descriptions are presented in Appendices. For example, it is quite likely that participant

Y41 who was first informed by her partner (over the phone) while in her office, heard the news again from her team leader and other colleagues. Similarly, participant O45 was quite likely to hear the news again on 12 September when checking in for a flight to Nimes (a salient event that one is less likely to confabulate), and participant O34 at a drycleaner's after first hearing the news in a school playground (see Appendix 2).

Memory descriptions of participant Y12, who was tested twice in September 2001, are particularly interesting as they provide a unique insight into how memory errors and distortions develop when people hear the news from multiple sources on the day of the public event. At initial test, on 12 September, this participant said he heard the news from a departmental secretary in the general office. At initial re-test on 27 September he repeated this account but also pointed out that he bumped into a colleague on a staircase immediately after leaving the general office. It is highly likely that the colleague also informed him of the attack[8] and consequently, in memory descriptions provided in 2003 and 2004, the participant believes that it was this colleague who he first heard the news from (see Appendix 3). The aforementioned fluctuations and memory errors are consistent with the Neisser and Harsch's (1992) wrong time slice hypothesis (see general discussion).

Relationship between consistency scores and background variables

The relationship between the probed recall consistency (WAS) and the background variables was examined in the entire sample, ignoring the factors of delay and number of tests in 2001 as they were not involved in any main or interaction effects in the ANOVAs reported earlier. None of the encoding variables (surprise, emotion, stress, etc.) were reliably correlated with the test-retest consistency scores in either 2003 or 2004 (see also Hirst et al., 2009). Table 3 shows correlations for post encoding variables obtained at the time of re-tests in 2003 and 2004. Only two sets of correlations with consistency scores were significant. There were small but positive correlations between the mean confidence

ratings in the accuracy of probed recall questions in 2003 and 2004 and the WAS in 2004. There were also small but negative correlations between rehearsal ratings in 2003 and the WAS in 2003 and 2004. Thus, consistency scores were lower the more participants reported remembering/thinking about their personal circumstances of hearing the news. Table 3 also shows that all the variables in 2003 were positively and reliably correlated with equivalent measures in 2004, and that vividness was positively related to confidence and rehearsal (remembering/thinking about personal circumstances).

General Discussion

The present study investigated the consistency of flashbulb memories of September 11 over long delay periods of 2- and 3-years as a function of a delay between the reception event and an initial test (1-2 days vs. 10-11 days), and the number of initial tests (one vs. two) in a sample of 124 adults from the general population. The aim of the study was to address several important questions in current flashbulb memory research including (1) the effects of long delays and repeated testing on flashbulb memories; (2) the Winningham et al. (2000) consolidation hypothesis; (3) the fluctuations in memory distortions and the Neisser and Harsch (1992) wrong time slice hypothesis; and (4) the association between background variables and the consistency of recall. Below, each of these questions is examined in light of the findings obtained in the present study.

Consistency of flashbulb memories over long delays

Although research on flashbulb memories has been growing steadily, there are only four studies that have examined the test-retest consistency over delays of 24 months and longer (Bohannon & Symons, 1992; Hirst et al., 2009; Neisser & Harsch, 1992; Schmolck et al., 2000).[9] Two of these studies, by Neisser and Harsch (1992) and Schmolck et al. (2000), revealed very high levels of distortion and inconsistency over delays of 32 to 34 months. As a result, Schmolck et al. (2000) suggested that while flashbulb memories are

retained relatively well within the first 12 months, major forgetting and distortions occur with delays between 15 and 32 months.

The results of the present study definitely do not support this claim, either for a test-retest delay of 23-24 months or a longer delay of 35-36 months (see also Hirst et al., 2009). Although the consistency of memories for September 11 dropped significantly from 2001 to 2003 and from 2001 to 2004 in half of the sample that was tested twice in 2001, the mean consistency scores (WAS) in 2003 and in 2004 were much higher than those reported by Neisser and Harsch (1992) and Schmolck et al. (2000). While 50% of participants in the Neisser and Harsch (1992) study obtained the lowest possible WAS of 0 to 2, only a small minority of our participants did so (7% in 2003 and 12% in 2004) (see Figure 2). Similarly, the percentage of memories classed as *major distortion* was markedly lower in our study (9% in 2003 and 7% in 2004) in comparison to 25% and 40% reported by Neisser and Harsch (1992) and Schmolck et al. (2000), respectively. It is also interesting, that only 2 participants, out of 124, were unable to recall their personal circumstances when asked to provide a memory description in 2003 and 2004. *Can't remember* responses were also very rare when participants were answering 5 questions in probed recall, with the majority of participants being able to provide substantive answers to each of the five questions (the percentages of substantive answers ranged from 94% to 100% across the five questions in 2003 and from 94% to 99% in 2004) (for similar results see Berntsen & Thomsen, 2005; Kvavilashvili et al., 2003; Tekcan & Peynircioglu, 2002).

One possible explanation for these discrepant findings concerns the number of times participants are tested. Participants in Neisser and Harsch (1992) and Schmolck et al. (2000) studies were tested only twice, once immediately after the reception event and then after a long delay. In our study, participants were tested either three times (in 2001, 2003 and 2004) or four times (twice in 2001 and once in 2003 and 2004). However, the results

of our study did show that an additional re-test in 2001 had no significant effect on the consistency scores or memory descriptions either in 2003 or in 2004. It is also unlikely that having a re-test in 2003 (23-24 months after the reception event) had any major beneficial effect on participants' scores in 2004 by eliminating a high percentage of distortions that would have otherwise developed between 23-24 months and 35-36 months. There are a growing number of studies which show that the number and timing of re-tests (e.g., whether the retest is conducted before or after the September 11 anniversary) does not affect the completeness of memory reports or their consistency (e.g., A.R.A Conway et al., 2009, Hirst, et al., 2009; Hornstein et al., 2003; Shapiro, 2006; see also Tizzard-Drover & Peterson, 2004). Therefore, it is more likely that the discrepancies are due to different public events being used in these studies (September 11 vs. Challenger explosion and O.J. Simpson's acquittal) or, in case of Schmolck et al. (2000), a possible confound from the second verdict of O.J. Simpson's civil trial, announced 16 months after the initial verdict.

The consolidation hypothesis (Winningham et al., 2000)

According to the consolidation hypothesis, those participants who are interviewed within the first few days after the event will exhibit poorer consistency across the test-retest sessions than those who are first tested after a longer delay from the original event. However, the results of the present investigation on flashbulb memories of September 11 do not provide any support for the consolidation hypothesis even though we had a sufficient power (.80) to detect small to medium effect sizes, and a power of .97 to detect medium effect sizes. Thus, participants who were initially tested on 12-13 September were no less consistent than those who were tested on 21-22 September either in terms of their memory descriptions (free recall) or in terms of their WAS for probed recall. The absence of an effect could be due to the fact that the consolidation of flashbulb memories occurs much faster (may be in the first 30-60 minutes from the reception event) than suggested by the consolidation hypothesis. For

example, new evidence from a recent animal study by Diamond and colleagues (cited in Diamond, Campbell, Park, Halonen & Zoladz, 2007) shows that rats who were briefly pre-exposed (for 2 minutes) to emotionally arousing stimulus (a cat), followed by a session with a minimal water maze training (only 4 trials), had strong spatial memory for the location of a hidden platform after a 24 hour delay interval. However, this effect was present only when the training occurred immediately after the emotional arousal but not when it occurred after 30 minutes, indicating that there is a fairly short time window in which emotional arousal improves the consolidation of temporally contiguous experiences.[10] If similar mechanisms are involved in the formation of flashbulb memories of reception context in humans then it is not surprising that we did not obtain any effect of delay of initial testing. It is possible that memory traces had already consolidated by the time we tested our participants on 12 and 13 September (i.e., 24 to 48 hours after the event).

Our results are also in conflict with the initial results of Winningham et al.'s (2000) study on flashbulb memories of the O.J. Simpson's acquittal, which supported the consolidation hypothesis. Other recent studies (mostly on flashbulb memories of September 11) that have specifically addressed this question have obtained inconsistent findings. For example, in the study of Schmidt (2004) with a test-retest delay of 2 months, consistency scores for September 11 were significantly higher for participants who were initially tested on 16-20 September than those who were tested on 12-13 September 2001, as the consolidation hypothesis would predict (for similar findings see Weaver & Krug, 2004). By contrast, in the study of Lee and Brown (2003) with a test-retest delay of 7 months, there was no difference in participants' consistency scores as a function of delay between the event (September 11) and an initial test (4-24 hours vs. 10 days). Bohannon (personal communication) also failed to obtain a significant effect of delay of initial testing in a study where participants were first tested either on 11 September (N=300), 12

September (N=100) or 1-2 weeks later (N=300) and were re-tested at intervals of 3 months, 1 year and 2 years (parts of this study, unrelated to consolidation hypothesis, are reported in Julian et al., 2009). Similar non-significant results were obtained by Coluccia et al. (2006) who initially tested their participants either 2, 18, 27 or 51 days after the Columbia shuttle explosion and then re-tested them after 5 and 12 months (see also results of Neisser et al., 1996 on a sample of students from Emory University).

It is interesting that in the majority of studies (two out of three) that did find support for the hypothesis (i.e., Schmidt, 2004; Winningham et al., 2000), only the probed recall test was used (i.e., participants only answered a set of specific questions about activity, location, etc.). However, in the four (out of five) studies that did not find any support for the hypothesis (our study, Bohannon, personal communication; Lee & Brown, 2003, Neisser et al., 1996) both free and probed recall tests were administered. It is therefore possible that discrepant results are due to this methodological difference between the studies. For example, having to provide a memory description immediately after the event may speed up/help the formation of a stable narrative account that will eliminate any differences that may otherwise exist between the participants who were tested immediately or after some time from the flashbulb event. Clearly, future studies need to be conducted in which having a free recall test (present vs. absent) and a delay between the event and an initial test (short vs. long) are contrasted orthogonally.

Fluctuations of memory descriptions and wrong time slice hypothesis

One of the most interesting set of findings obtained in the present study is that participants' memory descriptions were not static but changed across the re-tests in terms of wording, and the number and specificity of details/canonic categories mentioned (see Table 2). These fluctuations indicate that participants are reporting only a subset of information available to them at a particular time, perhaps the information that is most

activated or that they regard as the most important/pertinent at that particular time. This fits well with the self-memory system postulated by an influential theory of autobiographical memory (Conway, 2005; Conway & Pleydell & Pearce, 2000). Of particular importance is participants' memory descriptions classed as *major distortion* and their fluctuations across the two testing sessions in 2003 and 2004. Although the number of such memories was relatively small (see Figure 3) they provide valuable insights into the nature of flashbulb memories and a possible mechanism responsible for developing memory distortions.

The first important finding was that the number of *major distortions* (or even *minor distortions*) did not further increase in 2004. Similar findings were obtained by Neisser and Harsch (1992) who re-tested their participants one more time, 38-39 months after the Challenger explosion, and found virtually no further increase in the percentage of *major distortions* (28% as opposed to 25% at a 32-34 delay interval). It is interesting that in their study those participants who gave completely wrong accounts at the initial 32-34 month re-test, gave the same wrong accounts at the final 38-39 month re-test, i.e., their memories did not show any fluctuations across the testing sessions. In contrast, in our study, 36% of memories that were classed as *major distortions* in 2003 changed into *minor distortions*, *less specific* and even *same* memories in 2004 (see Table 2). Results did not change when we pooled major distortions, minor distortions and can't remember responses into a category of inconsistent memories and less specific, more specific and same responses into a category of consistent memories: 31% of memories that were *inconsistent* in 2003 changed into *consistent* memories in 2004. Similar fluctuations in flashbulb memories across the 11- and 35-month re-tests were recently reported by Hirst et al. (2009). In their study, approximately 28% of memories of September 11 coded as inconsistent at initial re-test were re-coded as consistent at subsequent re-test.

Neisser and Harsch (1992) explained the formation of major distortions by participants remembering another episode of hearing the news and incorrectly assuming it was the very first time when they heard it. According to the wrong time slice hypothesis, distortions are not complete confabulations or false memories but rather refer to real episodes of hearing the news again either on the same day or a subsequent day; hence the high levels of reported vividness and confidence in the accuracy of memories classed as distortions (see Brewer, 1992; Weaver & Krug, 2004).

Consider the *major distortions* reported by participants O45 and O34 in 2003 (see Appendix 2). It is quite likely that after hearing the news in the car (O45) and school playground (O34), they heard the news again in the airport the next day and at the dry cleaner's a quarter of an hour later, respectively, and for some reason these subsequent occasions were perhaps more emotional/informative and replaced the initial memories of hearing the news in the car and at the playground. Indeed at the final retest in summer 2004 they stuck to their distorted memory descriptions provided in 2003.

Further evidence for the wrong time slice hypothesis comes from those participants whose memories fluctuated between different episodes across the two re-testing sessions. For example, participant Y31 reported initially hearing the news at work from a friend but in 2003 said she heard it from her husband when she went home for a lunch break. However, in 2004 she again said she heard the news first at work (but this time from a customer) and then at home during a lunch break. Similarly, participant O57 gave a different account of hearing the news in 2003 but in 2004 she reverted back to the exactly same account she provided in September 2001 (see Appendix 4).

Another interesting example comes from participant Y44, a social worker, who initially said he heard the news from a friend while at work (when getting ready for a community visit) but at the re-test in 2003 provided a detailed account of how he heard the

news on a scooter on his way to a community visit. However, in 2004 he explicitly said he had two memories of hearing the news, one in his office and one on his scooter.

The development of these errors and distortions in participants' memories is not surprising giving the gradual nature of the development of the news story on 11 September 2001. Lots of people initially heard that there was a plane crash and then heard additional news as the events unfolded during the afternoon of that day. Given the gradual nature of the news about September 11, and extensive media coverage, it is quite surprising that the percentage of 'major distortions' was not, in fact, higher than we found (9% in 2003 and 7% in 2004). One can even argue that because of this methodological confound, the consistency of memories of unique personal events (e.g., being told about the death of a parent or an accident) should be much higher than the consistency of flashbulb memories of public events (see Pillemer, 2009, for a more detailed discussion of this point).

Additional support for the wrong time slice hypothesis comes from correlational analyses. Thus, small but reliable negative correlations were obtained between the ratings of rehearsal in 2003 (remembering and/or thinking about one's personal circumstances of hearing the news) and probed recall consistency scores in 2003 and 2004: the higher rates of rehearsal were related to lower levels of consistency (see Table 3). This is not the first study to report negative correlations between rehearsal and consistency. Although the majority of studies have reported either non-significant (e.g., Davidson & Glisky, 2002; Hirst et al., 2009; Pillemer, 1984; Schmolck et al., 2000 for a delay of 32 months) or positive correlations between the two (Curci & Luminet, 2006; Julian et al., 2009; Otani et al., 2001; Schmolck et al., 2000 for a delay of 15 months), Bohannon and Symons (1992) found a negative relationship between rehearsal and consistency scores for flashbulb memories of Challenger explosion (see Talarico & Rubin, 2009 for a further discussion of this issue). Interestingly, this correlation was significant only in participants who reported

being upset by the news but not in those who were not upset. A possible explanation of the negative correlation is that over the delay interval, people may think about several different occasions when they heard the news on 11 September (not only the first one), and this will negatively affect their consistency scores for the first reception event.

Although results concerning memory fluctuations and the wrong time slice hypothesis are interesting, they are based on the qualitative analysis of a relatively small number of memory distortions. One interesting way of assessing the wrong time slice hypothesis in more controlled manner is to ask participants at initial test to recall all those situations on the day of the event when they were informed about the news (not only the first occasion). Having these initial reports will allow one to assess whether a memory coded as major distortion at re-test is a confabulation or a real (but not the first) episode of hearing the news. Clearly, this is an interesting avenue for future research in this area.

Associations between consistency scores and background variables

Although our results did not support the idea that rehearsal plays an important role in post encoding consolidation processes, the correlations between the consistency scores and encoding variables such as surprise, emotion/stress and importance were also non significant. All of these variables have been implicated in the formation of flashbulb memories by various models of flashbulb memories (for a review see Luminet, 2009). However, empirical results on this issue are mixed (see Talarico & Rubin, 2009). For example, in those studies that do report reliable correlations between the consistency and self-rated emotional intensity, the correlations are fairly small, usually between .20 and .30 (Conway et al., 1994; Curci & Luminet, 2009a; Pillemer, 1984), and several studies, even with fairly large participant numbers, have failed to obtain significant correlations (Curci & Luminet, 2006; Davidson & Glisky, 2002; Hirst et al., 2009; Neisser & Harsch, 1992; Neisser et al., 1996; Otani et al., 2005; Shapiro, 2006; Smith et al., 2003).

One possible reason for low and/or absent correlations is ceiling effects in participants' ratings (e.g., ratings of surprise and national importance in the present study). However, ratings of emotion and stress were clearly not at ceiling. It was our impression that some British participants tended to downplay their emotional reactions. This is in line with the study of Kvavilashvili et al. (2003) in which British participants' had as vivid and detailed flashbulb memories of September 11 as Italian participants but their ratings of emotion were reliably lower than those of Italians (see Curci & Luminet, 2006 for similar results with Japanese and the US participants). To obtain more valid measures of self-rated emotion (and other variables) some recent studies have asked people to provide several different ratings tapping into different aspects of emotion and using composite scores instead of single measures (e.g., Curci & Luminet, 2009a; Talarico & Rubin, 2003; 2007).

Conclusions

In summary, several important findings emerged from the present study that have both theoretical and methodological implications for flashbulb memory research. First, it appears that data collected within the first 10-11 days will be no less reliable than the data obtained within 1-2 days of the event. Methodologically, this is a useful finding since it indicates that it is not essential to collect data immediately after the reception event. Second, results showed that flashbulb memories are neither immune to forgetting nor static formations that remain identical across testing sessions, as originally proposed by Brown and Kulik's (1977) 'Now Print' theory. This is because there was both a reliable drop in the consistency scores over the first two years and fluctuations in memory descriptions (in terms of categories they were assigned to) (*cf*. Hirst et al., 2009). In addition, a close examination of major distortions in 2003 and 2004 provided further evidence in support of the wrong time hypothesis, which explains the development of distortions by hearing the news from multiple sources on the day of the flashbulb event.

However, contrary to Neisser and Harsch (1992) and Schmolck et al. (2000), our results also showed remarkably good retention of flashbulb memories over a long delay of 35-36 months. Only 2% of participants could not remember the reception event in 2004, and the ratings of vividness were as high after three years as in September 2001 (see also Kvavilashvili et al., 2003; Tekcan et al, 2003; Talarico & Rubin, 2007). Most importantly, flashbulb memories were more consistent than in the Neisser and Harsch (1992) and Schmolk et al. (2000) studies, both in terms of the mean WAS and the percentage of major distortions. In this respect, our findings are in line with several recent studies on flashbulb memories of September 11 that have also demonstrated a fairly good consistency with delays ranging from 1 to 2 years (A.R.A. Conway et al., 2009; Curci &Luminet, 2006; Julian et al., 2009; Shapiro, 2006; Tekcan et al., 2003). Our findings not only replicate but significantly extend the results of these studies by showing that there was no further increase in major distortions from 2003 (delay of two years) to 2004 (delay of three years). Clearly, flashbulb memories of some public events are robust, even if far from perfect.

Finally, our results provide no evidence in support of the idea that flashbulb memories are primarily maintained by post encoding rehearsal (see Neisser & Harsch. 1992; Winningham et al., 2000). The additional re-test administered to half of the sample two weeks after their initial test in September 2001, did not have any beneficial effect on subsequent consistency scores in 2003 and 2004 (for similar findings see Hirst et al., 2009; Hornstein et al., 2003; Shapiro, 2006; Tizzard-Drover & Peterson, 2004). There was also no positive relationship between these scores and self-reported rehearsal of the reception event in the delay interval (indeed some of the correlations were negative and reliable). The results also did not support the consolidation hypothesis which suggests that the post-encoding consolidation of memory trace may take several days or weeks after the event: participants who were first tested 10-11 days after the event (allowing sufficient time for

consolidation to occur) did not have better consistency scores than those who were tested 1-2 days after the event. This finding, together with the absence of any positive effects of repeated testing and rehearsal suggests that flashbulb memories are formed automatically and consolidated fairly soon, possibly within an hour, from the reception event.

Additional support for this idea comes from recent laboratory studies of emotional memories which show that emotionally arousing words and pictures facilitate the automatic encoding of contextual information which is not mediated by rehearsal or other processes such as distinctiveness (Anderson, Wais, & Gabrieli, 2006; Mackay & Ahmetzanov, 2005; MacKay, Shafto, Taylor, Marian, Abrams & Dyer, 2004). For example, in a study of Anderson et al. (2006), emotionally arousing pictures enhanced participants' long-term memory for neutral pictures presented 4 seconds earlier (a process similar to remembering irrelevant contextual details of the news event in flashbulb memories), while no such enhancement was present for highly distinctive but non emotional pictures. This enhanced recall disappeared when the delay between the neutral and emotional stimuli was increased to 9 seconds. The authors conclude that "emotionally arousing events result in a direct neurobiological enhancement of memory consolidation, independently of attention to and elaboration with the to-be-remembered event" (p. 1602).

It is of course unclear whether similar (cognitive and brain) mechanisms are involved in the recall of emotionally arousing stimuli (words, pictures, films) in the laboratory and flashbulb memories in everyday life (e.g., Sharot, Martorella, Delgado, & Phelps, 2007). However, studying flashbulb memory-like consolidation processes in the laboratory can provide interesting and valuable information. As Curci and Luminet (2009b) recently put it, flashbulb memories "can not be considered as an isolated phenomenon, but they need to be analysed with respect to broader research questions concerning the general functioning of individual's cognitive system" (p. 274).

References

Anderson, A. K., Wais, P. E., & Gabrieli, D. E. (2006). Emotion enhances remembrance neutral events past. *Proceedings of National Academy of Sciences, 103*, 1599-1604.

Berntsen, D., & Thomsen, D. K. (2005). Personal memories for remote historical events: Accuracy and clarity of flashbulb memories related to World War II. *Journal of Experimental Psychology: General*, 134, 242-257.

Bohannon, J. N., & Symons, L. V. (1992). Flashbulb memories: Confidence, consistency, and quantity. In E. Winograd & U. Neisser (Eds.), *Affect and accuracy in recall: Studies of "flashbulb" memories* (pp. 65-91). Cambridge: Cambridge University Press.

Brewer, W. F. (1988). Memory for randomly sampled autobiographical events. In Neisser and E. Winograd (Eds.), *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (pp. 21-90). New York: Cambridge University Press.

Brewer, W. F. (1992). The theoretical and empirical status of the flashbulb memory hypothesis. In E. Winograd & U. Neisser (Eds.), *Affect and accuracy in recall: Studies of "flashbulb" memories* (pp. 274-305). Cambridge: Cambridge University Press.

Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition, 5*, 73-99.

Christianson, S. -Å. (1989). Flashbulb memories: Special, but not so special. *Memory & Cognition, 17*, 435-443.

Christianson, S. -Å., & Engelberg, E. (1999). Memory and emotional consistency: The MS Estonia ferry disaster. *Memory, 7*, 471-482.

Cohen, G., Conway, M. A., & Maylor, E. A. (1994). Flashbulb memories in older adults. *Psychology and Aging, 9*, 454-463.

Collucia, E., Bianco, C., & Brandimonte, M. (2006). Dissociating veridicality, consistency and confidence in autobiographical and event memories for the Columbia shuttle disaster. *Memory, 14*, 452-470.

Conway, A. R. A., Skitka, L. J., Hemmerich, J. A., & Kershaw, T. C. (2009). Flashbulb memories for September 11, 2001. *Applied Cognitive Psychology, 23,* 605-623.

Conway, M. A. (1995). *Flashbulb memories*. Hillsdale: Lawrence Erlbaum Associates.

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language*, *53*, 597-628.

Conway, M. A., Anderson, S. J., Larsen, S., Donnely, C. M., McDaniel, M. A., McClelland, A. G. R, Rawles, R. E., & Logie, R. H. (1994). The formation of flashbulb memories. *Memory & Cognition, 22,* 326-343.

Conway, M. A. & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self memory system. *Psychological Review, 107*, 261-288.

Cubelli, R., & Della Sala, S. (2008). Flashbulb memories: Special but not iconic. *Cortex, 44,* 908-909.

Curci, A., & Luminet, O. (2006). Follow-up of a cross-national comparison on flashbulb and event memory for September 11th attacks. *Memory, 14*, 329-344.

Curci, A., & Luminet, O. (2009a). Flashbulb memories for expected events: A test of the emotional-integrative model. *Applied Cognitive Psychology, 23*, 98-114.

Curci, A., & Luminet, O. (2009b). General Conclusions. In O. Luminet & A. Curci (Eds.), *Flashbulb memories: New issues and new perspectives* (pp. 269-276). Hove and New York: Psychology Press.

Curci, A., Luminet, O., Finkenauer, C., & Gisle, L. (2001). Flashbulb memories in social groups: A comparative test-retest study of the memory of French President Mitterand's death in a French and a Belgian group. *Memory, 9*, 81-101.

Davidson, P. S. R., Cook, S. P. & Glisky, E. L. (2006). Flashbulb memories for September 11th can be preserved in older adults. *Aging, Neuropsychology, and Cognition, 13,* 196-206.

Davidson, P. S. R., & Glisky, E. L. (2002). Is flashbulb memory a special instance of source memory? Evidence from older adults. *Memory, 10*, 99-111.

Diamond, D. M., Campbell, A. M., Park, c. R., Halonen, J., & Zoladz, P. R. (2007). The temporal dynamics model of emotional memory processing: A synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson Law. *Neural Plasticity, 2007*, 1-33.

Ferré Romeu, P. (2006). Memories of the terrorist attacks of September 11, 2001: A study of the consistency and phenomenal characteristics of flashbulb memories. *The Spanish Journal of Psychology, 9*, 52-60.

Hirst, W., Phelps, E. A., Buckner, R. L., Budson, A. E., Cuc, A., et al. (2009). Long-term memory for the terrorist attack of September 11: Flashbulb Memories, event memories, and the factors that influence their retention. *Journal of Experimental Psychology: General, 138*, 161-176.

Horn, D. B. (2001). Confounding the effects of delay and interference on memory distortion: Commentary on Schmolck, Buffalo, and Squire. *Psychological Science, 12*, 180-181.

Hornstein, S. L., Brown, A. S., & Mulligan, N. W. (2003). Long-term flashbulb memory for learning of Princess Diana's death. *Memory, 11*, 293-306.

Julian, M., Bohannon, J. N., & Aue, W. (2009). Measures of flashbulb memory: Are elaborate memories consistently accurate? In O. Luminet & A. Curci (Eds.), *Flashbulb memories: New issues and new perspectives* (pp. 99-122). Hove and New York: Psychology Press.

Kvavilashvili, L., Mirani, J., Schlagman, S., & Kornbrot, D. E. (2003). Comparing flashbulb memories of September 11 and the death of Princess Diana: Effects of time delays and nationality. *Applied Cognitive Psychology, 17*, 1017-1031.

Kvavilashvili, K., Mirani, J., Erskine, J. A. K., Schlagman, S., & Kornbrot, D. E. (2009). Effects of age on phenomenology and consistency of flashbulb memories of September 11 and a staged control event. Manuscript submitted for publication.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Larsen, S. F. (1992). Potential flashbulbs: Memories of ordinary news as the baseline. In E. Winograd, & U. Neisser (Eds.), *Affect and accuracy in recall: Studies of "flashbulb memories"* (pp. 32-63). Cambridge: Cambridge University Press.

Lee, P. J., & Brown , N. R. (2003). Delay related changes in personal memories for September 11, 2001. *Applied Cognitive Psychology, 17*, 1007-1015.

Livingston, R.B. (1967). Brain circuitry relating to complex behavior. In G. C. Quarton, T. Melnechuk, & F. O. Schmitt (Eds.), *The neurosciences: A study program* (pp. 499-514). New York: Rockefeller University Press.

Luminet, O. (2009). Models for the formation of flashbulb memories. In O. Luminet & A. Curci (Eds.), *Flashbulb memories: New issues and new perspectives* (pp. 51-76). Hove and New York: Psychology Press.

Luminet, O., & Curci, A. (2009) (Eds.). *Flashbulb memories: New issues and new perspectives*. Hove and New York: Psychology Press.

Luminet, O., Curci, A., Marsh, E. J., Wessel, I., Constantin, T., Gencoz, F., & Yogo, M. (2004). The cognitive, emotional, and social impacts of the September 11 attacks: Group differences in memory for the reception context and the determinants of flashbulb memory. *The Journal of General Psychology, 131*, 197-224.

MacKay, D. G., & Ahmetzanov, M. V. (2005). Emotion, memory, and attention in the taboo Stroop paradigm: An experimental analogue of flashbulb memories. *Psychological Science, 16,* 25-32.

MacKay, D. G., Shafto, M., Taylor, J. K., Marian, D. E., Abrams, L., & Dyer, J. (2004). Relations between emotion, memory and attention: Evidence from taboo Stroop, lexical decision, and immediate memory tasks. *Memory & Cognition, 32*, 474-488.

McCloskey, M. (1992). Special versus ordinary memory mechanisms in the genesis of flashbulb memories. In E. Winograd, & U. Neisser (Eds.), *Affect and accuracy in recall: Studies of "flashbulb memories"* (pp. 227-235). Cambridge: Cambridge University Press.

McCloskey, M., Wible, C. G., & Cohen, N. J. (1988). Is there a special flashbulb memory mechanism? *Journal of Experimental Psychology: General, 117,* 171-181.

Neisser, U. (1982). Snapshots or benchmarks? In U. Neisser (Ed.), *Memory observed: Remembering in natural contexts* (pp. 43-48). New York: W. H. Freeman.

Neisser, U., & Harsch, N. (1992). Phantom flashbulbs: False recollections of hearing the news about Challenger. In E. Winograd, & U. Neisser (Eds.), *Affect and accuracy in recall: Studies of "flashbulb memories"* (pp. 9-31). Cambridge: Cambridge University Press.

Neisser, U., Winograd, E., Bergman, E. T., Schreiber, C. A., Palmer, S. E., & Weldon, M. S. (1996). Remembering the earthquake: Direct experience vs. hearing the news. *Memory, 4,* 337-357.

Niedzwienska, A. (2004). Metamemory knowledge and the accuracy of flashbulb memories. *Memory, 12,* 603-613.

Otani, H., Kusumi, T., Kato, K., Matsuda, K., Kern, R. P., Widner, W. Jr., & Ohta, N. (2005). Remembering a nuclear accident in Japan: Did it trigger flashbulb memories? *Memory, 13*, 6-20.

Pezdek, K. (2003a). Event memory and the autobiographical memory for the events of September 11, 2001. *Applied Cognitive psychology, 17*, 1033-1045.

Pezdek, K. (Ed.) (2003b). Memory and cognition for the events of September 11, 2001. *Applied Cognitive Psychology, 17* (A Special Issue).

Pillemer, D. B. (1984). Flashbulb memories of the assassination attempt on President Reagan. *Cognition, 16*, 63-80.

Pillemer, D. B. (2009). "Hearing the news" versus "being there": Comparing flashbulb memories and recall of first-hand experiences. In O. Luminet & A. Curci (Eds.), *Flashbulb memories: New issues and new perspectives* (pp. 125-140). Hove and New York: Psychology Press.

Rubin, D. C. (1992). Constraints on memory. In E. Winograd, & U. Neisser (Eds.), *Affect and accuracy in recall: Studies of "flashbulb memories"* (pp. 265-273). Cambridge: Cambridge University Press.

Schmidt, S. R. (2004). Autobiographical memories for the September 11th attacks: Reconstructive errors and emotional impairment of memory. *Memory & Cognition, 32*, 443-454.

Schmolck, H., Buffalo, E. A., & Squire, L. R. (2000). Memory distortions develop over time: Recollections of the O.J. Simpson's trial verdict after 15 and 32 months. *Psychological Science, 11*, 39-45.

Shapiro, L. R. (2006). Remembering September 11th: The role of the retention interval and rehearsal on flashbulb and event memory. *Memory*, *14*, 129-147.

Sharot, T., Martorella, E., A., Delgado, M. R., & Phelps, E. A. (2007). How personal experience modulates the neural circuity of memories of September 11, *Proceedings of National Academy of Sciences, 104*, 389-394.

Smith, M. C., Bibi, U., & Sheard, D. E. (2003). Evidence for the differential impact of time and emotion on personal and event memories for September11, 2001. *Applied Cognitive Psychology, 17,* 1047-1055.

Talarico, J. M., & Rubin, D. C. (2003). Confidence, not consistency characterizes flashbulb memories. *Psychological Science, 14*, 455-461.

Talarico, J. M., & Rubin, D. C. (2007). Flashbulb memories are special after all; in phenomenology, not accuracy. *Applied Cognitive Psychology*, *21*, 557-578.

Talarico, J. M., & Rubin, D. C. (2009). Flashbulb memories result from ordinary memory processes and extraordinary event characteristics. In O. Luminet & A. Curci (Eds.), *Flashbulb memories: New issues and new perspectives* (pp. 80-97). Hove and New York: Psychology Press.

Tekcan, A. I., Ece, B., Gülgöz, S., & Er, N. (2003). Autobiographical and event memory for 9/11: Changes across one year. *Applied Cognitive Psychology, 17,* 1057-1066.

Tekcan, A. I., & Peynircioglu, Z. F. (2002). Effects of age on flashbulb memories. *Psychology and Aging, 17,* 416-422.

Tizzard-Drover, T., & Peterson, C. (2004). The influence of an early interview on long-term recall: A comparative analysis. *Applied Cognitive Psychology, 18*, 727-743.

Weaver, C. A. (1993). Do you need a "flash" to form a flashbulb memory? *Journal of Experimental Psychology: General, 122*, 39-46.

Weaver, C. A., & Krug, K.S. (2004). Consolidation-like effects in flashbulb memories: Evidence from September 11, 2001. *American Journal of Psychology, 117*, 517-530.

Winningham, R. G., Hyman, I. E., & Dinnel, D. L. (2000). Flashbulb memories? The effects of when the initial memory report was obtained. *Memory, 8,* 209-216.

Winograd, E. (1992). Introduction. In E. Winograd and U. Neisser (Eds.), *Affect and accuracy in recall: Studies in "flashbulb" memories* (pp. 1-5), New York: Cambridge University Press.

Winograd, E., & Neisser, U. (1992) (Eds.), *Affect and accuracy in recall: Studies of "flashbulb memories"*. Cambridge: Cambridge University Press.

Winograd, E., & Killinger, W.A. Jr. (1983). Relating age at encoding in early childhood to adult recall: Development of flashbulb memories. *Journal of Experimental Psychology: General, 112,* 413-422.

Walters, G., & Goudsmit, J. (2005). Flashbulb and event memory of September 11, 2001: Consistency, confidence and age effects. *Psychological Reports, 96*, 605-619.

Wright, D. B. (1993). Recall of the Hillsborough disaster over time: Systematic biases of "flashbulb" memories. *Applied Cognitive Psychology, 7,* 129-138.

Wright, D. B., & Gaskell, G. D. (1995). Flashbulb memories: Conceptual and methodological issues. *Memory, 3*, 67-80.

Yarmey, a. D., & Bull, M. P. (1978). Where were you when President Kennedy was assassinated? *Bulletin of the Psychonomic Society, 11*, 133-135.

Footnotes

[1] Other possible ways of assessing the special status of flashbulb memories is to examine the role of several encoding (e.g., surprise, emotion, perceived importance) and post-encoding (e.g., rehearsal) variables on the consistency of flashbulb memories, or comparing the consistency of flashbulb memories to that of some control (personal or non-personal) event.

[2] In all other test-retest studies of flashbulb memory, only one initial test is obtained and consistency is assessed by comparing responses at this initial test with those of subsequent re-test(s).

[3] Since this factor did not have any effect on the dependent variables, results will be reported on the entire sample.

[4] Due to experimenter error, ratings of importance were not obtained at initial interviews in September 2001.

[5] Details of this coding scheme can be obtained from the first author upon request.

[6] The data was pooled because factors of initial delay and number of tests in 2001 did not produce any effects either for 2003 (p=.82 and p=.37, respectively) or for 2004 re-tests (p=.50 and p=.69, respectively).

[7] The emergence of *more specific* memories was unexpected and raises an important question about the accuracy and nature of retention of flashbulb memories. It has been suggested that they represent memory distortions and confabulations developing over time, as in ordinary autobiographical memories (McCloskey et al., 1988; Schmidt, 2004; Smith et al., 2003). However, others have assumed that the amount of available detail may stay more or less same over the delay but, at any given moment, the current goals of "working self" will determine the number and/or specificity of details

recalled. Therefore, these researchers would not necessarily regard more specific memories at the re-test as distortions (e.g., Conway et al., 1994; Conway & Pleydell-Pearce, 2000; Niedzwienska, 2004; Winningham et al., 2000).

[8] Indeed, this colleague, who was also a participant in the study, reported that he heard the news from a friend who called him while he was working in his office, at about the same time that Y12 was in the building, and that he left the office to tell the others about the news and watch the TV upstairs. It is therefore, highly likely that he also told the news to Y12 who he met on the stairs. However, if he did not, then Y12 committed a source misattribution error at re-test, instead of the wrong time slice error. Although this possibility cannot be ruled out entirely, it is more likely that Y12 and the colleague did exchange the news when they met on the staircase.

[9] Although studies by A.R.A Conway et al. (2009), Julian, Bohannon, and Aue (2009, and Shapiro (2006) have also used a delay of 2 years, participants in their studies were additionally re-tested at earlier intervals.

[10] Coincidentally, in another study described by Diamond et al. (2007), rats also developed a strong, extinction resistant, fear of context in which they had a brief encounter with a cat.

Table 1

Mean Ratings of Background Variables at Initial Test in September 2001 and Subsequent Re-Tests in Summer 2003 and Summer 2004 (Standard Deviations in Brackets). Right Hand Columns Present Results of One-Way ANOVAs on These Means (F and P Values and Effect sizes). All Ratings Were Made on 10-Point Rating Scales (1=Not at All, 10=Extremely).

	Year of Testing					
	2001	2003	2004	F value (2,242)*	p value	Partial η^2
Surprise	8.65 (2.01)	9.07 (1.64)	9.00 (1.64)	3.27	.05	.027
Emotion	4.90 (2.59)	6.85 (2.51)	6.29 (2.56)	36.04	<.00001	.23
Stress	6.09 (2.57)	6.16 (2.68)	6.33 (2.37)	.75	.47	.00
Importance (personal)	–	6.03 (2.91)	5.80 (2.48)	1.40	.24	.01
Importance (national)	–	8.75 (1.31)	8.75 (1.29)	.005	.94	.00
Vividness of reception event	8.39 (1.93)	8.30 (1.72)	8.26 (1.70)	.28	.76	.00
Rehearsal of September 11	7.36 (1.87)	5.75 (1.78)	2.75 (1.73)	302.80	<.00001	.71
Rehearsal of Reception event	–	3.09 (1.78)	4.65 (2.21)	49.99	<.00001	.29
Confidence in free recall	–	8.46 (1.55)	8.67 (1.48)	1.94	.17	.02
Confidence in probed recall	–	8.99 (.90)	9.11 (.77)	2.53	.11	.02

* Degrees of freedom for variables that were obtained only in 2003 and 2004 were 1, 121

Note. Bonferroni correction was applied for post hoc comparisons between the means.

Table 2

Frequency of Memory Descriptions, Relative to Descriptions in 2001, in Each of the Six Response Categories in 2004 as a Function of Response Categories in 2003 (Row Percentages in Brackets).

		Memory Descriptions in 2004						
		Can't remember	Major distortion	Minor distortion	Less specific	More specific	Same	Total
Memory Descriptions in 2003	Can't remember	1 (50%)	0	0	1 (50%)	0	0	2 (100%)
	Major distortion	0	7 (64%)	2 (18%)	1 (9%)	0	1 (9%)	11 (100%)
	Minor distortion	0	2 (7%)	17 (59%)	8 (27%)	2 (7%)	0	29 (100%)
	Less specific	1 (4%)	0	7 (26%)	15 (55%)	3 (11%)	1 (4%)	27 (100%)
	More specific	0	0	6 (13%)	13 (28%)	22 (48%)	5 (11%)	46 (100%)
	Same	0	0	1 (11%)	6 (67%)	1 (11%)	1 (11%)	9 (100%)
	Total	2 (2%)	9 (7%)	33 (27%)	44 (35%)	28 (23%)	8 (6%)	124 (100%)

Table 3

Pearson Product Moment Correlations Between Weighted Attribute Scores in 2003 (WAS2003) and 2004 (WAS2004) and Self-Rated Post-Encoding Variables Obtained in 2003 and 2004. All Ratings Were Made on 10-Point Rating Scales (1=Not at All, 10=Extremely).

	WAS 2003	WAS 2004	Confid. 2003	Confid. 2004	Rehear. 2003	Rehear. 2004	Vivid 2003	Vivid 2004
WAS 2003	1.00							
WAS 2004	.73****	1.00						
Confidence 2003	.14	.18*	1.00					
Confidence 2004	.36***	.33***	.54****	1.00				
Rehearsal 2003	-.25**	-.25**	.16	.05	1.00			
Rehearsal 2004	-.05	.05	.004	.09	.27**	1.00		
Vividness 2003	.17*	.17	.49****	.32**	.29**	.02	1.00	
Vividness 2004	.12	.11	.37***	.47****	.25**	.27**	.53****	1.00

*p<.05; **p<.01; ***p<.001; ****<.00001

Figure Captions

Figure 1. Mean Consistency Scores (WAS) as a Function of Delay From Initial Testing (Short vs. Long), Number of Tests in 2001 (1-Test vs. 2-Tests) and Time of Re-Test (2001 vs. 2003 vs. 2004). Data for 2001 Re-Test is Available only for 65 Participants Who Were Tested Twice in 2001.

Figure 2. Frequency Distribution of Weighted Attribute Scores in Our Study (in 2003 and 2004) and in Neisser and Harsch (1992).

Figure 3. Percentages of Memory Descriptions Coded as 'Can't Remember', 'Major Distortion, 'Minor Distortion' 'Less Specific', 'More Specific', and ''Same' as a Function of Year of Re-Test (2003 vs. 2004).

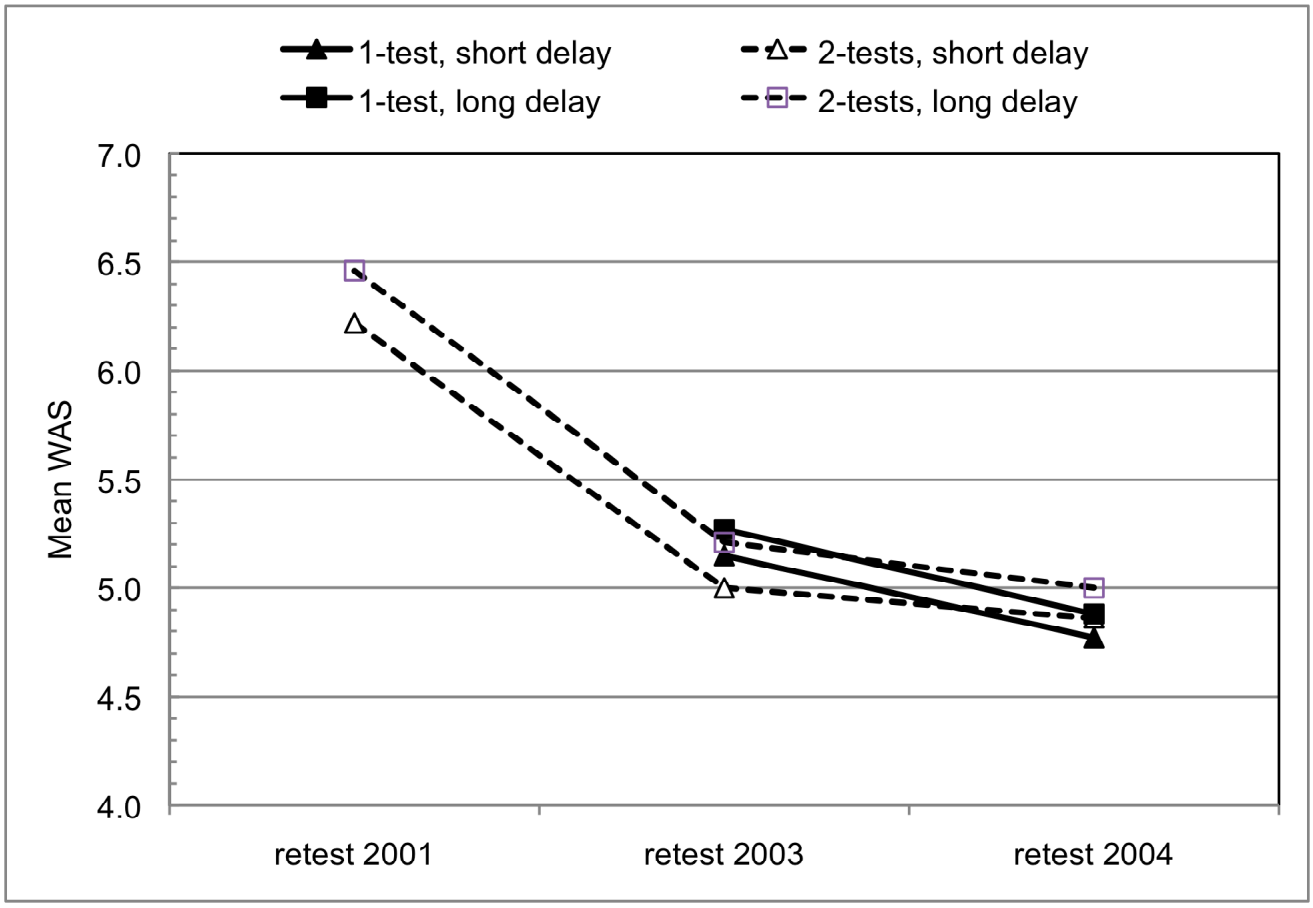
1-test, short delay
2-tests, short delay
1-test, long delay
2-tests, long delay
7.0
6.5
6.0
5.5
5.0
4.5
4.0
Mean WAS
retest 2001
retest 2003
retest 2004

Figure 1

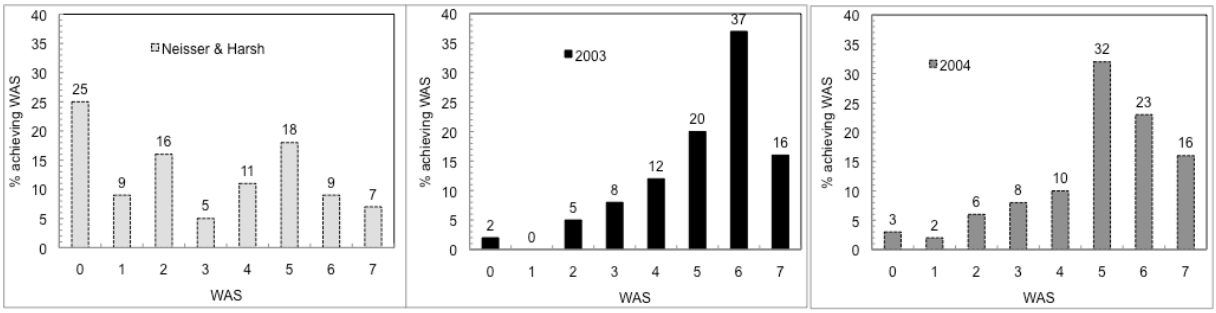
Neisser & Harsh
% achieving WAS
WAS
25
9
16
5
11
18
9
7
2003
2
0
5
8
12
20
37
16
2004
3
2
6
8
10
32
23
16

Figure 2

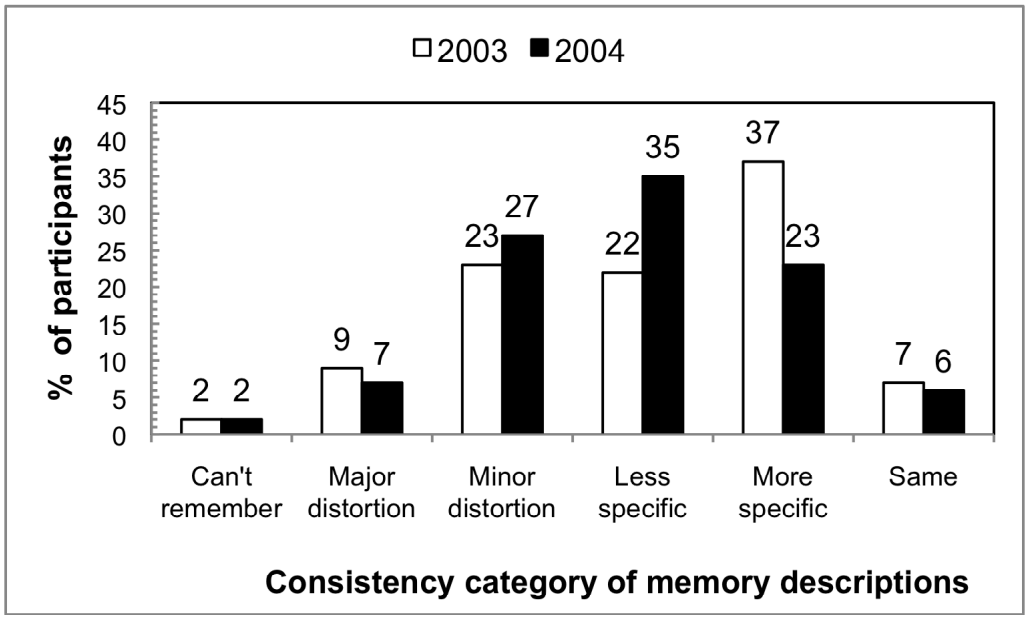
2003
2004
% of participants
45
40
35
30
25
20
15
10
5
0
2
2
9
7
23
27
22
35
37
23
7
6
Can't remember
Major distortion
Minor distortion
Less specific
More specific
Same
Consistency category of memory descriptions

Figure 3

Appendix 1.

Coding Scheme for Assessing Consistency in Probed Recall (after Neisser & Harsch, 1992)

Type of response at re-test and the assigned score			
	Can't remember or major distortion 0	Less specific or minor distortion 1	Same or More specific 2
Time	Can't remember *Major distortion:* Initially 'morning' then 'evening' If more than 3 hours different within a specific day time period, e.g., initially '7 am' then '11:30 am'	*Less specific:* 'At 3 pm' then 'in afternoon' or 'between 3 and 4' *Minor distortion:* 'At lunchtime' then 'between 2-3'	*Same:* If within 15 –20 minutes of originally stated time *More specific*: 'Between 4 and 4:30' then 'at 4:15' 'In the morning' then 'between 10 and11am'
Location	Can't remember *Major distortion:* 'At home' then 'at work'	*Less specific:* 'At desk at work' then 'at work' *Minor distortion:* 'in kitchen' then 'in lounge' (but still at home!)	*Same:* If same place as before *More specific:* 'At home' then 'at home in kitchen'
Activity	Can't remember *Major distortion*: 'Was driving' then 'was shopping'	*Less specific*: "Watching 'Friends' then 'watching TV' *Minor distortion:* 'was writing emails' then 'typing a document on PC'	Same: If same activity as before *More specific:* 'I was working' then ' I was working on the computer'
Others Present	Can't remember *Major distortion*: 'My husband' then 'my neighbour' or 'alone'	*Less specific:* 'My friend Mark' then 'my friend' *Minor distortion:* 'My friends Jo, Tom and Sandra' then 'my friends Sandra and Jo'	*Same:* If same person(s) as before *More specific*: 'My friend' then 'my friend John'
Source	Can't remember *Major distortion:* 'Heard it on radio' then 'my friend told me'	*Less specific:* 'On radio 4' then 'on radio' *Minor distortion*: 'My colleague Sam told me' then 'my colleague Mark told me'	*Same*: If same source as before *More specific:* 'on radio' then 'on radio 4'

Appendix 2

Examples of Memory Descriptions Classed as Major Distortions in 2003.

Participant Y41

Initially tested on 13 September, 2001: "At work, and Jeff (partner) phoned me and asked if I'd heard and then he told me that there had been a terrorist attack on the World Trade Center and the Pentagon. And it took me a few seconds to realize where it had happened, because I could not think where the W.T.C was".

Re-test on 12 August, 2003: "I got a call from my team leader (Wendy) saying have you heard the news. At the same time someone came round the corner and also told me New York was under attack. After speaking to her on the phone I went round to look at the TV in the office where I saw the 2nd plane go into the 2nd tower".

(*Confidence rating – 7; vividness – 7*).

Participant 045

Initially tested on 21 September, 2001: "I was returning from a shopping trip and I put on the car radio and heard something on the news and it was very early and the news story was just breaking and they interrupted Radio 4… As I arrived home I just sat and watched it".

Re-test on 27 July 2003: "We were standing in a queue (Judy and I) waiting to check in for a flight to Nimes the day after – the 12th of September (a Wednesday or Thursday). We were quite horrorstruck and surprised. We had been packing up in the night before and we were subject when we checked in to the most vigorous and thorough luggage and body search to Nimes"

(*Confidence rating - 9, vividness - 9*).

Participant 034

Initially tested on 12 September, 2001: Complete shock, I went to pick my granddaughter from school at 2:45 pm and two ladies in the playground told me about this horrible event".

Re-test on 18 August 2003: "I was in Hatfield in a dry cleaners picking up some dry cleaning and the chap there told me. It was about 3 o'clock in the afternoon". (*Confidence rating - 7, vividness -10*).

Appendix 3

Memory Descriptions of Participant Y12 at Initial Test and Subsequent Re-Tests in 2001, 2003 and 2004.

Initially tested on 12 September, 2001:"I was in the departmental office and Margaret said have you heard the news, there's been a terrorist attack in America, they are watching it on the TV upstairs".

Re-tested on 27 September, 2001: "I heard about it in the general office and Margaret said something's going on in America, there's been some aircraft crashing and I went upstairs to look at the TV and bumped into Cameron on the stairs".

Re-tested on 1 September, 2003: "I was in the stairwell on my way to Psychology, CP Snow. I think Cameron came out and said 'have you heard about this thing in NY an aeroplane has crashed into the WTC.'. I said 'Is it a terrorist attack?'. I'm pretty sure it was before the 2nd plane crashed and it was about lunchtime. I then went upstairs to the top floor and watched a live feed and saw the 2nd plane crash. I stayed there for 1.5 hours-maybe more" *(Confidence rating - 8, vividness - 7).*

Final re-test in August 2004: "I was going into the Psychology department and I bumped into Cameron who said 'A plane has flown into WTC'. Then I went upstairs and watched it on TV in 2H255" *(Confidence rating - 8, vividness - 6).*

Appendix 4

Memory Descriptions of Two Participants Whose Memories Were Classed as 'Major Distortions' in 2003 but Were Re-Classed into Different Categories in 2004.

Participant Y31

Initially tested on 12 September, 2001: "Emptying till at work and work friend (Olivia) told me and I thought she was joking to frighten me because I was going to fly abroad tomorrow. She told me a plane had crashed".

Re-tested on 15 August, 2003: "I was at work, came home for lunch. Nick (husband) said it had just happened on the TV – planes had crashed into the towers. I went back to work and everyone was talking about it. And the 13th of that month I was going on my first long flight so it made me very anxious" *(Confidence rating - 6, vividness - 8).*

Final re-test on 12 August, 2004: " I was at work it was a customer that said something had gone off in America. It wasn't until I came home at lunch that it was all being relayed through the TV" *(Confidence rating - 5, vividness - 8).*

Participant 057

Initially tested on 20 September, 2001: "We have our house for sale and my husband was at Estate Agents and a chap there who is selling our house had told him about it. And my husband came straight home, told me and we put the TV on".

Re-tested on 19 August, 2003: "I was out at the time – I came home and my husband told me about it" *(Confidence rating -10, vividness -10).*

Final re-test on 27 July, 2004: "I was in the lounge at home. We'd just put the house on the market. My husband came home from the Estate Agents and said there had been bombings in New York, so we switched the TV on and saw it" *(Confidence rating -10, vividness –10).*

Author Note

Lia Kvavilashvili, Jennifer Mirani, and Diana Kornbrot, School of Psychology, University of Hertfordshire, UK.

Simone Schlagman is now at Inter-Research Science Centre, Oldendorf/Luhe, Germany

Kerry Foley, School of Psychology, University of Leicester, UK.

Research presented in this paper was supported by a research grant to Lia Kvavilashvili and Diana Kornbrot from the Economic and Social Research Council (UK), and a British Academy/Leverhulme Trust (UK) senior research fellowship awarded to Lia Kvavilashvili. We are grateful to Laura Fisher, James Erskine, David Wellsted, and Rebecca Macham for conducting telephone interviews at various stages of the project, and to Jenna Garner for helping us out with coding the data. Portions of research described in this paper were presented at the British Psychological Society Cognitive Section Conference in September 2005 (University of Leeds), Psychonomic Meeting in November 2005 (Toronto), Bial Foundation meeting in April 2006 (Porto) and the 4th International Conference on Memory in July, 2006 (University of New South Wales, Sydney).

Correspondence concerning this article should be addressed to Lia Kvavilashvili, School of Psychology, University of Hertfordshire, College Lane, Hatfield, Herts, AL10 9AB, UK. Email: *L.Kvavilashvili@herts.ac.uk*