APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **17**: 1007–1015 (2003)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/acp.982

Delay Related Changes in Personal Memories for September 11, 2001

PETER JAMES LEE* and NORMAN R. BROWN

University of Alberta, Canada

SUMMARY

This study examined delay related changes of people's recollections for 11th September 2001. 1481 participants were surveyed 4–24 hours or 10 days after the event. 142 participants were re-tested in April, 2002. Test-retest consistency was low after seven months (66.5%). Word counts for open ended descriptions revealed that people wrote significantly more contextual information 10 days after the event than respondents had on 11th or 12th September although no difference was found for retest participants 7 months later. Ratings for emotional reaction decreased monotonically over time. These results suggest early indexing may be a critical factor if the amount of information reported, type of information reported, or level of affect is a research issue. However, test-retest consistency was not influenced by the ten day delay in indexing. Copyright © 2003 John Wiley & Sons, Ltd.

Important, surprising, and impactful public events sometimes produce *flashbulb memories*. These are vivid and long lived recollections of the personal circumstances associated with learning about such events (Brown & Kulik, 1977; Conway et al., 1994). Subsequently, researchers have used a test-retest design to assess the accuracy and stability of these memories over fairly long delays (e.g. Neisser & Harsch, 1992; Neisser et al., 1996; Weaver, 1993). This line of work has demonstrated that memory reports collected months or years after the flash-bulb-eliciting event sometimes differ from ones collected days or weeks after the event.

The test-retest method rests on the assumptions that the initial memory report provides a veridical account of the target reception event, and that differences between the initial report (*indexing*) and follow-up reports provide evidence for the inaccuracy or instability of flashbulb memories. The problem with this approach is that there is inevitably some delay between the reception event and the initial test.[1] In principle, this delay could be problematic as post-event mechanisms (e.g., narrativation, schematization, socially-mediated rehearsal) might distort or overwrite memory for the initial reception event, and post-event experiences concerning the flash-bulb eliciting event might interfere with people's ability to access the original reception-event memory, even after relatively short delays. In either case, observed test-retest differences would be difficult to interpret. Of

*Correspondence to: Peter J. Lee, Department of Psychology, University of Alberta, Edmonton, AB T6G 2E9, Canada. E-mail: pjlee@ualberta.ca

[1]In a recent review of 15 test-retest studies of flashbulb memories, Winnigham et al. (2000) found that delay between the reception event and indexing ranged from 1 day to 1 month with a median of 3 days.

course, it is possible that reception events for emotionally powerful news events are encoded so well, and remembered with such clarity, that these memories are immune to both the potentially distorting post-event processes and the interfering effects of related event memories, at least for a while. If so, it is reasonable to treat test-retest consistency as an accuracy measure.

This study addressed the problem of delay of indexing. We present a study that indexed participants starting four hours after the World Trade Center / Pentagon attacks, and compare subsequent test-retest reports with people who were first indexed ten days after the event. We focus on three aspects we believe will interest flashbulb memory researchers; the amount of information people report over time, changes in levels of affect, and test-retest consistency.

Only one study to date has compared people's memory hours after an event with reports collected a few days later (Winningham, Hyman, & Dinnel, 2000), the target event was the verdict in the O. J. Simpson homicide trial. These researchers measured consistency over time using the test-retest method. They also measured the amount of information that people recalled by counting the number of propositions used by participants to answer open-ended questions about their personal circumstances at the time of the event. Results indicated that respondents tested on the day of the event were less consistent at retest than respondents who were first tested a week later. The word counts also revealed that people reported less information eight weeks after the event, although there were no differences in word counts at initial testing, i.e. there was a monotonic decrease in reported information over time. Winningham et al. (2000) proposed a model they call *consolidation* to explain the differences in consistency and word counts. This position suggests that emotive memories are pruned of extraneous information over relatively short periods of time, but the central details become more robustly encoded. The consolidation of event information indicates that rehearsal and other post-event experiences may have influenced people's memory for the target event.

In summary, an important question facing flashbulb memory researchers is the significance they should place on early indexing, and the effects of delay (if any). The differences in consistency reported by Winningham et al. (2000) suggest that early indexing has important implications, especially when the research aim is to accurately measure test-retest consistency. However, the verdict in the O. J. Simpson murder trial was not particularly surprising and had been the subject of considerable media speculation for over 16 months. The verdict was also the culminating incident in a narrative, and can be considered as a consequence rather than a uniquely surprising event. In this paper we attempt to address some of the concerns about the importance of early indexing by comparing the effects of delay on reports elicited 4–24 hours after the 11th September attack on the World Trade Center / Pentagon, with reports collected ten days later. A subset of these reports are then compared with information collected at retest seven months later.

METHOD

Four factors permitted a rapid response to the events on 11th September 2001. A comprehensive flashbulb questionnaire, originally designed by M. A. Conway (unpublished), was immediately available. The two-hour time difference between New York City and Edmonton, Canada, and the time of the attacks (08.45 EDT), meant that researchers could begin preparing materials before many participants became aware of the event.

Ethical approval was rapidly expedited after three hours of scrutiny, consultation and revision. Lastly, numerous instructors made large classes available for immediate testing. In this section we only provide details salient to the results presented here.

Participants

1481 undergraduates from the University of Alberta participated. Their median age was 19 years. One group of participants, Wave 1, was tested between 4 hours and 24 hours after the initial attack on the World Trade Center ($n = 697$). A second group, Wave 2, was tested ten days after the event on 21st September 2003 ($n = 784$). Finally, 142 participants were invited back in the first two weeks of April, 2002 (Wave 1, $n = 72$, Wave 2, $n = 70$).[2]

Participants were selected and assigned to Wave on an opportunity basis. Although opportunity sampling is a weak sampling strategy, the problem with non-random sampling was overcome by sample size. The number of introductory psychology students tested was 54% ($n = 1180$) of the department of psychology's participant pool. These students would have had better than an even chance of being selected from the pool if a random sampling technique had been employed. There was no reason to believe that the classes assigned to Waves 1 & 2 on the basis of opportunity (e.g. a chance meeting with a course instructor in an elevator) were either different or systematically biased. The remaining 301 students were undertaking either introductory social psychology (Wave 1) or introductory anthropology (Wave 2).

Across all 1481 participants, there were no significant between wave differences ($p > 0.10$) in age, gender or nationality, indicating these groups were fairly homogenous.

Procedure

Participants were seated and waiting for the start of undergraduate classes when they were asked to take part in a study about the terrorist attacks in the U.S. Participants were told that the research was concerned with people's memory and reactions to important public events. They were informed that their cooperation was entirely voluntary and that no incentive would be offered.

Waves 1 and 2 received identical questionnaires. The first question asked respondents to provide a short open-ended description (approximately half a page) about the circumstances in which they first learned of the attack on the World Trade Center and the Pentagon.

Respondents were then asked more specific questions about whom they were with (*people*), their whereabouts (*location*), and what they were doing (*activity*). The people questions began with a simple fixed choice (Yes/No) response to the question 'was anyone with you when you first heard the news'. If 'yes', they were then asked to provide the names of each individual (or group e.g. 104 psych. class).The location question asked participants to provide a brief description (one sentence) identifying their location on hearing the news. The activity question was also a brief description asking about what the respondent was doing on hearing the news.

Participants then answered four emotional response questions using a 5-point rating scale (1 = no emotion, 5 = intense arousal). They were asked to rate how surprised, sad,

[2]The retested respondents were a subset of those who agreed to be contacted, and do not represent a low (10%) proportion of agreement. Participants who returned at times other than April 2002 are not being reported in this article.

shocked and upset they felt when they first heard about the attacks. Finally, respondents were asked if they might be contacted for a follow up study. If so, the same questionnaire was administered at retest.

RESULTS AND DISCUSSION

Word counts

A measure of the amount of information people reported was made by calculating word counts for the initial open-ended descriptive question. A boxplot analysis of the word counts for the descriptive question indicated that 45 respondents had written nothing, or were extreme values (>2 times the interquartile range from their group median). These respondents were eliminated from all subsequent analyses.

A comparison between Wave 1 and Wave 2 word counts at initial testing was conducted using an independent samples *t*-test. Wave 2 participants wrote more words than Wave 1 participants. The mean word count for Wave 2 was 56.96 ($SD = 26.15$) compared with a mean of 50.59 words ($SD = 21.70$) for Wave 1 respondents, $t(1434) = 4.99$, $p < 0.01$. The seventy-two Wave 1 participants who were subsequently invited back to be retested demonstrated a similar pattern as their cohorts; mean word count = 51.72 ($SD = 19.31$). However, the seventy retested participants from Wave 2 wrote more information than the unretested participants from their cohort; mean word count = 63.23 ($SD = 22.53$). Figure 1 shows the mean word counts for the 142 retested participants from Wave 1 and Wave 2.

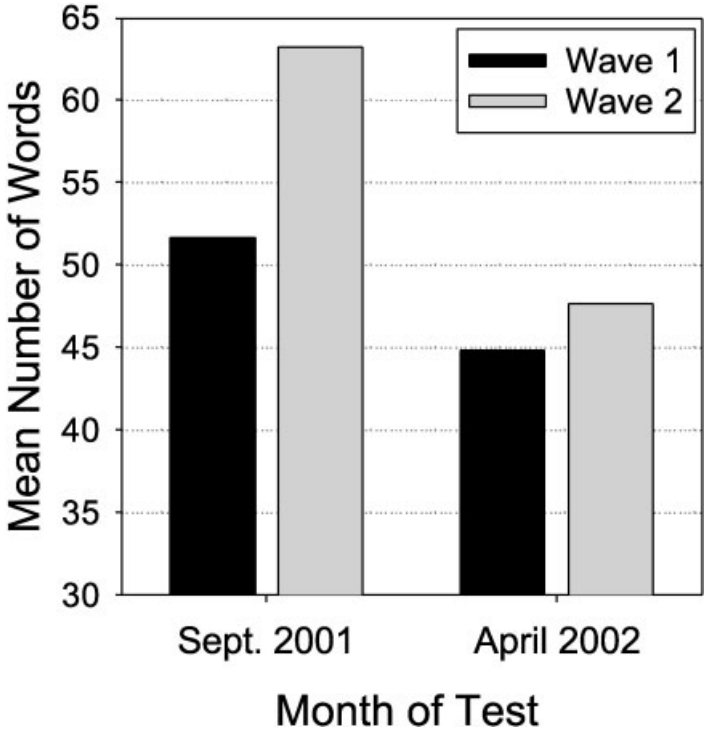

Figure 1. Mean word count as a function of Wave and test

A 2 (Wave) × 2 (test-retest) mixed factor ANOVA (Wave as a between subjects factor, and test-retest as within subjects) was conducted on the word counts for the 142 participants tested in September 2001 and again in April 2002. There was a main effect between test-retest, $F(1, 140) = 35.65$, $MSE = 248.89$, $p < 0.01$, indicating that participants wrote significantly less information in April, 2002, than they had previously in September 2001. There was also a significant main effect of Wave, $F(1, 140) = 7.39$, $MSE = 496.07$, $p < 0.01$ and a Wave by test-retest interaction, $F(1, 140) = 5.32$, $MSE = 248.89$, $p < 0.05$. A post-hoc comparison showed there was no difference between Waves at retest, $t(140) = -0.977$, $p > 0.10$.

One explanation for this dissociation is that Wave 2 retested participants are unrepresentative of Wave 2 as a whole, and were biasing Wave 2 word counts upwards. A 2 (Wave) × 2 (retested-unretested) ANOVA was conducted on the word counts at indexing, revealing a main effect of Wave, $F(1, 1432) = 16.58$, $MSE = 582.62$, $p < 0.01$, and a marginal effect of retest-unretested, $F(1, 1432) = 3.66$, $MSE = 582.62$, $p < 0.06$. More importantly, there was no significant interaction, $F(1, 1432) = 1.74$, $MSE = 582.62$, $p > 0.10$. A comparison between retested and non-retested participants showed that Wave 2 retested participants did indeed write more than their non-retested cohort, $t(760) = -2.11$, $p < 0.05$, while no re-sampling bias was found for Wave 1, $t(672) = -0.47$, $p > 0.10$. Nonetheless, a comparison between the un-retested participants from Waves 1 & 2 revealed that Wave 2 still provided more written information ($M = 56.33$ words, $SD = 26.43$) than Wave 1 ($M = 50.45$ words, $SD = 21.98$), $t(1289) = -4.34$, $p < 0.01$.

Although no difference was found in the proportion of people who agreed to be contacted for a follow up study versus those who actually turned up for retesting (in either Wave), the absolute number of respondents indicating a willingness to return when asked at indexing decreased from 73.3% at Wave 1 to 47.4% at Wave 2. The difference in sampling bias between Waves is readily explained by Wave 2 retested participants being more enthusiastic than Wave 1. Nevertheless, the data from the unretested participants still indicate that people wrote more ten days after the event than those tested within 24 hours.

What remains unclear is why respondents provided more written information on 21st September than people tested within 24 hours. The Yerkes-Dodson's law (Yerkes & Dodson, 1908) is an affective state approach that might explain this difference, based on the curvilinear relationship between arousal and performance. Perhaps participants tested within 24 hours may have experienced high enough levels of emotional arousal to negatively affect the amount of information they were capable of retrieving. Easterbrook's (1959) cue utilization hypothesis would imply that higher levels of affect negatively influenced Wave 1 participant's ability to access cue information compared to Wave 2 participants.

An alternative explanation is that respondents experience a form of facilitated retrieval after a period of rehearsal. This position assumes that respondents limit the amount of time and cognitive effort they are willing to expend on retrieving contextual details and that the association between cues and memory becomes stronger over time. This approach is supported by the facilitated cue retrieval model proposed by Ratcliff and McKoon (1988), which suggests that familiarity (which we assume to be enhanced by rehearsal over a period of days) results in more rapid and easily retrieved memories. It is possible that Wave 2 respondents were quicker at accessing the same information as Wave 1 respondents and were subsequently able to invest more time, and additional resources, retrieving detailed information.

Our explanations are speculative and are not meant to be exhaustive. They do, however, indicate that memory reports for highly emotional events might be influenced not only by the quality of initial encoding but also by post-event factors such as rehearsal and contemporary affect. Still, it appears that delay in the indexing of highly salient events affects the amount of information people report. Respondents willing to participate in follow-up studies tended to report more information when asked ten days after an event than unwilling participants or people asked soon afterwards. This increase in reported information is problematic, because it introduces the possibility that delays may bias the sample (given that the choice to complete follow-up studies is largely out of the researchers control). However, our retested participants from Wave 1 appear to be an unbiased sample of their cohort. These data suggest that compliance with our appeal to be retested is higher and more likely to result in a representative sample if people are indexed very soon after an event.

Emotional arousal

A measure of overall emotional arousal was calculated by summing over the four affective states; surprised, sad, shocked, and upset. The revised scale ranged from 0 (no emotional reaction across any measure) to 20 (very intense emotions for all measures). A comparison between Waves showed that Wave 1 participants reported being more highly aroused on hearing the news than Wave 2, $t(1431) = 2.55$, $p < 0.05$.

Participant's affect scores at test and retest were entered into a 2 (Wave) $\times$ 2 (test-retest) mixed factor ANOVA. Figure 2 shows the mean arousal ratings by Wave and test. There was a main effect of test showing that people's overall level of arousal had reliably

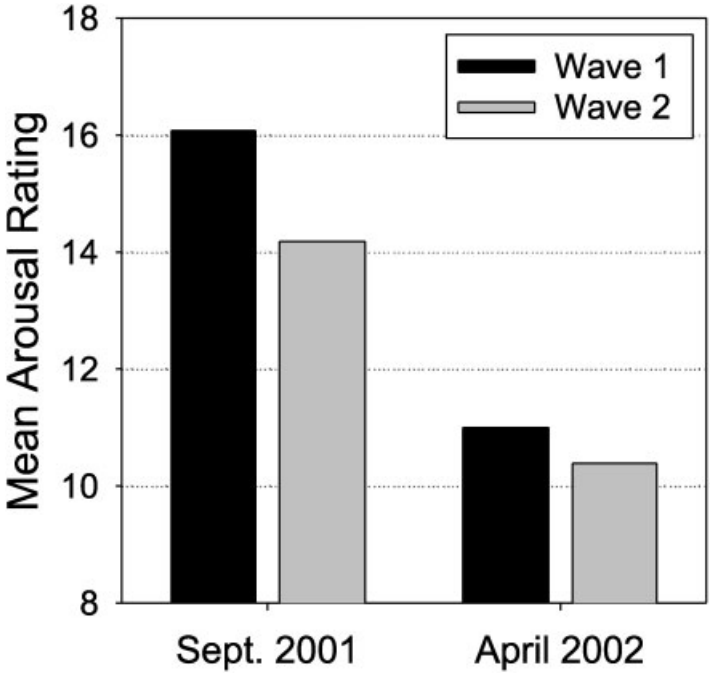

Figure 2. Mean level of emotional arousal as a function of Wave and test

decreased, as might be expected after 7 months, $F(1, 140) = 206.56$, $MSE = 6.71$, $p < 0.01$. There was also a main effect of Wave, $F(1, 140) = 7.63$, $MSE = 14.36$, $p < 0.01$. A significant interaction showed that Wave 2 respondents showed less emotional arousal than Wave 1 respondents at initial testing, but no differences between groups at retest, $F(1, 140) = 4.75$, $MSE = 6.71$, $p < 0.05$ (see Figure 2).

These data indicate that emotional arousal begins to decrease rapidly after an event, even one as momentous as 11th September. Once again, the data suggest that researchers interested in the relationship between affect and flashbulb memories, may want to consider indexing as soon as possible.

Consistency

The method used to score consistency was a 3-point scale developed by Neisser and Harsch (1992), and similar to the method used by other researchers (e.g. Cohen, Conway, & Maylor, 1994; Winningham et al., 2000). We used this scheme to analyze the people, location and activity questions. Completely consistent reports were scored as 2. Partially consistent reports were scored as 1. For example, 'I was in the shower when my mom called me' at test versus 'I was drying my hair when my mom called me' at retest show a high degree of convergence but are not exactly alike. Completely inconsistent reports received a score of 0 and were defined as reports whose content was entirely different at retest (e.g. 'I was driving to campus'; 'I was in class').

Two participants failing to report adequate information were omitted from this analysis. Table 1 shows the percentage of reports rated as consistent, partially consistent, and inconsistent by person, location and activity. The overall level of consistency is poor considering the delay between test and retest. However, it is also the case that the number of completely inconsistent reports is also low. This difference is explained by the proportion of partially consistent reports, which suggests finer grained information first reported at test was omitted at retest.

Using a technique similar to Winningham et al. (2000), a composite score was constructed by summing across all levels of consistency, and the proportion of flashbulb memories was then estimated using these scores. This new scale ranged from 0 (completely inconsistent across people, location and activity) to 6 (completely consistent across the three measures). Participants exhibiting flashbulb memories were classified as those people whose composite consistency scores were 5 or above, while participants with scores below 5 were categorized as not having flashbulb memories. Using this method, 66.5% of retested respondents exhibited flashbulb memories. Considering the nature of the target event, and the short delay between test and retest, this figure is low and inconsistent

Table 1. Percentage of scores at each level of consistency for people, location, and activity

	Target memory					
	People		Location		Activity	
Consistency	Wave 1	Wave 2	Wave 1	Wave 2	Wave 1	Wave 2
Consistent	68	68	87	90	72	68
Partial	16	17	11	3	16	13
Inconsistent	16	15	1	7	13	19

with a mechanism that encodes information subsequently available over long periods of time. No differences, however, were found between Waves at any level of consistency, context type, or the number of flashbulb memories, χ^2 (2, N = 142) = 0.6, $p > 0.10$.

Because test-retest consistency has been a major concern for researchers in this area, these data provide some reassurance that moderate delays prior to testing do not unduly affect consistency measures at retest. The discrepancy between these findings and Winningham et al. (2000) may be due to the nature of the target event. Perhaps a period of mnemonic consolidation takes place quite soon after an event if the event is highly salient and emotive, but not particularly surprising or consequential. Memories for 11th September may not have undergone this process of consolidation if information about people's personal context had been sufficiently encoded on hearing the news. This latter position suggests that people's memories are largely unaffected over the short term, but information is gradually lost over extended periods of time (i.e. a ceiling effect at the time of the event, with a monotonic decrease in consistency over time).

CONCLUSIONS

In this study we initially tested one group of people shortly after 11th September, and another group ten days later, to investigate delay related changes in memory for highly emotional events. Our aim was to clarify some of the methodological concerns about the potential effects of delayed indexing on measures of performance at retest. In doing so we also presented an unusual finding about the amount that people report after short delays and speculated about the causes of this phenomenon. We provided converging evidence that argues against the special encoding mechanism for flashbulb memories(cf. McClosky, Wible, & Cohen, 1988; Neisser & Harsch, 1992).

The importance of early indexing has been a question asked by many researchers investigating flashbulb memories. The results presented here are prescriptive, but depend on the aim of the investigation. If the focus is to examine the amount of contextual information that people recall, the details they report, or relationship between affect and memory then early indexing (e.g. within 24 hours) might be the preferred research strategy. On the other hand, consistency between test and retest does not appear to be sensitive to delays as long as ten days prior to test. We believe that the relative stability in consistency over short delays is encouraging, and should increase our confidence when comparing studies where results rely on accurate measures of consistency.

ACKNOWLEDGEMENTS

This research was supported by an NSERC grant awarded to the second author. The authors would like to acknowledge Courtney Bryden for help throughout this project. We would also like to thank Peter Dixon for providing invaluable advice during a rapid but comprehensive ethical review.

REFERENCES

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

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

Conway, M. A., Anderson, S. J., Larsen, S. F., Donnelly, 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.

Easterbrook, J. A. (1959). The effect of emotion on the utilization and organization of behaviour. *Psychological Review, 66*, 183–201.

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

Neisser, U., & Harsch, N. (1992). Phantom flashbulbs: false recollections of hearing the news about Challenger. In E. Winograd, & U. Neisser (Eds.), *Affect & consistency 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.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review, 95*, 385–408.

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

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.

Yerkes, R., & Dodson, J. (1908). The relation of strength of stimulus to rapidity of habit-information. *Journal of Comparative Neurology and Psychology, 18*, 459–482.

