CrossMark

Officer-involved shootings: testing the effect of question timing on memory accuracy for stressful events

Louise E. Porter [1,2] • Justin Ready [1,2] • Geoffrey P. Alpert [1,3]

Abstract

Objectives This randomised controlled trial tested the effect of immediate versus delayed, as well as repeated, questioning on memory retrieval regarding details of officer-involved shootings.

Methods A sample of 87 police officers experienced "active armed offender" training scenarios followed by a memory questionnaire testing their recall and recognition of details of the scenario. Officers were randomly assigned to one of two experimental groups. Group 1 completed the questionnaire immediately after the training scenario and again 2 days later, while group 2 completed the survey 2 days later only.

Results Linear mixed models showed that delayed questioning negatively impacted officers' recognition of correct details of the incident, but only for details not directly associated with threat stimuli. This pattern was supported by the free recall data. Further, officers who were questioned immediately did not experience the same memory decay over time, showing that early questioning can aid memory retention. Officers were also better at recognising correct threat-relevant details (compared to non-threat relevant) that were visual, but the reverse was true for auditory details.

Conclusions We found no support for the position that delay improves officers' memory. However, rather than finding universal memory decay over time, the interaction between timing and type of details being recalled revealed a more nuanced picture. This knowledge assists understanding of the limitations of information provided by officers under different interviewing conditions.

Keywords Police · Memory · Stress · Interviewing · Shooting

✉ Louise E. Porter
  l.porter@griffith.edu.au

[1] Griffith Criminology Institute, Griffith University, Mount Gravatt, Australia

[2] School of Criminology and Criminal Justice, Griffith University, Southport, QLD 4222, Australia

[3] Department of Criminology and Criminal Justice, University of South Carolina, Columbia, SC, USA

⌖ Springer

Introduction

When to interview an officer after an incident of lethal force, such as a shooting, is an under-researched question that has been the subject of recent attention in both academia and applied contexts (Grady et al. 2016). An officer-involved shooting (OIS) is an emotional, traumatic and stressful event that can impact the officer in a variety of ways, including their perceptions of what occurred (see Klinger 2007; Klinger and Brunson 2009). While officers who are injured or suffering noticeable trauma should not be immediately subjected to the additional stress of an interview, there are clearly investigative priorities, as well as public pressures, which necessitate an officer's account of the incident. A debated issue is when to interview those officers who are capable or able to be interviewed, in order to ensure the most accurate account of the incident.

Current thinking is split between an immediate and delayed interview. Arguments for the latter tend to focus on the benefits of delay for "emotional decompression and memory consolidation" (Lewinski et al. 2016, p. 64), stating that "Delay enhances an officer's ability to more accurately and completely respond to questions" (Lewinski, quoted in Force Science News n.d.). In contrast, the case for an immediate interview is rooted in the effects of memory decay and contamination. Unfortunately, no authoritative study has been conducted to develop evidence-based policy recommendations concerning the appropriate timing of interviews after critical incidents. Much of our understanding of memory comes from research on eyewitness testimony, with few studies on police officers' memory for incidents to which they have actively responded. Our study explores the accuracy of police officers' memory of an armed offender live action role-play scenario. Specifically, we test the impact of timing of questioning on accuracy of memory for details that are both central to understanding officers' decision-making and relevant to an investigation of the incident. We aim to increase understanding of police officer memory of stressful events and add empirical evidence to the debate over when to interview officers after a shooting in order to improve accuracy.

Effects of emotional state on memory and cognition

One argument for a delayed officer interview after a critical incident is based on the impact of stress on memory. As Grady et al. (2016, p. 248) state, "agencies believe that if an officer gives a report while under high levels of stress, the report will be less accurate and complete than if the report was given later under lower levels of stress". Eyewitness memory research suggests that stress impairs memory accuracy and eyewitness identification ability (see Deffenbacher et al. 2004, for a meta-analysis). Hope (2016) further discusses stress and anxiety as detrimental to police officers' memory for critical incidents. However, the focus of these studies is on the experience of stress at the time of memory *encoding*, while, as Grady et al. (2016) note, the rationale for a delayed interview after a stressful event is based on the relationship between memory and stress experienced at the time of memory *retrieval*. Much of the research conducted on stress and memory, while helpful in understanding deficiencies in officers' memory for stressful events, is less applicable to the debate over when officers should be interviewed to optimise the accuracy of recall.

It is likely, though, that officers may continue to experience the effects of stress after involvement in a stressful event (Hope 2016). Less is known about the effects of such

residual anxiety on either officers' subsequent event-specific memory or their general ability to process information in an interview setting. As Wolchover et al. (2014, p. 267) discuss, the enduring biological stress response from the incident could affect officers' "ability to focus on producing such an account at that stage (and hence, the subsequent reliability of that account)". This could be due to an impact of stress on general cognitive capability. For example, in a study of Taser exposure, White et al. (2015) showed that being Tasered temporarily impaired subsequent cognitive functioning, as measured by verbal learning ability (although it did not affect other forms of cognition). While Taser has a direct physiological impact, White et al. (2015, p. 606) also found psychological effects on participants' subjective state, which "raise the possibility that emotional factors following TASER exposure are important and may affect test performance". Thus, the subjective experience of stress or anxiety may influence an officers' cognitive capability to provide an account in an interview, at least in the short term. Our study seeks to explore the role of cognition in the relationship between timing of questioning and accuracy of officers' memory for shooting events.

Effects of delay versus early questioning on memory and cognition

A further argument for a delayed interview is based upon the beneficial role of sleep in memory consolidation. In support of this view, Geiselman (2010) showed rest positively correlated with eyewitness recall and suggested officers may not be well-rested if interviewed immediately after an incident. There is some evidence from laboratory-based studies that sleep may play an important role in memory, particularly regarding emotional memory processing (Genzel et al. 2015; Stickgold and Walker 2013). However, while a period of delay involving sleep can improve the amount of details recalled, it can also enhance the likelihood of false recall (false memories); particularly, memories can be less accurate after sleep (Payne et al. 2009).

In a recent review of research on memory and cognition, Grady et al. (2016) reported evidence that memory decays over time, with more opportunity for contamination the longer the delay period. The studies reviewed also showed that early questioning can aid memory retention over time. This is consistent with the literature on the "testing effect" (Roediger III and Karpicke 2006), which shows that prior testing of target material improves long-term memory retention beyond simply studying the material. In the eyewitness context, Burke et al. (1992) conducted two experiments to explore the effect of a 1–2-week delay on students' memory accuracy for a neutral versus emotionally arousing story (depicted on slides), using a multiple choice test. For emotionally arousing stories, immediate recall was more accurate than delayed recall over all, although the effects of delay were not stable across both their experiments. Further, they showed that participants tested both immediately and again 1 week later showed no difference in their memory over time. Burke et al. (1992, p. 285) suggested that early questioning "locked" in memory. However, they also noted that participants may have felt compelled to provide responses consistent with their initial answers, rather than actually remembering details from the event itself. Indeed, in the context of officer-involved shootings, the consequences of officers providing inconsistent answers about the incident across multiple interviews can be serious, including reflecting poorly on the officers' perceived credibility. Further, while early questioning may improve

memory for the subject material, this locking in of memory for certain details may be at the expense of memory for other details that were not the subject of initial testing. For example, studies of retrieval-induced forgetting have demonstrated more decay for non-prompted items than might naturally occur (Murayama et al. 2014). Thus, early questioning could reduce memory accuracy for some details.

While Grady et al. (2016) concluded that questioning early after an event would be more likely to produce accurate recall, this was predominantly based on general studies of memory and cognition (such as those outlined above) rather than applied research in the policing context. Grady et al. (2016, p. 249) found that "few studies meet the criteria to generalize to the practice of delayed reporting following an officer-involved shooting". They cite only two that compare the memory of sworn officers immediately after a (simulated) shooting versus a delay period; Beehr et al. (2004) tested a long delay of 12 weeks, while Alpert et al. (2012) explored a 3-day delay but did not test differences in recall for statistical significance. Both studies showed a tendency for less accurate recall after a delay, but that immediate questioning could improve longer-term memory for the event. A more recent study by Hartman et al. (2017) found officer's memory 4 to 10 weeks after involvement in a critical incident training scenario had not changed significantly (in terms of the total number of details or proportion of accurate details) from their memory immediately after the scenario. Our study seeks to add to this literature by empirically testing the effect of immediate versus 2-day delayed questioning on officers' memory, as well as the effects of repeated questioning.

The moderating effects of salience and sensory type on memory for shootings

Possibly confounding the debate over interview timing are factors related to the content of the material to be remembered. A number of studies of eyewitness memory using laboratory-based methods have shown that "central" information is more accurately recalled than "peripheral" information, particularly for highly emotional events (Christianson et al. 1990). The experiments by Burke et al. (1992, p. 287), noted earlier, showed that recognition of background detail suffered in the arousal condition (the emotional story), but was improved "for details that happened to be associated spatially with the event's center". However, what is considered central versus peripheral can depend on the context (Powell et al. 2009).

Shapiro (2006) explored students' recall of a video of a bike theft and defined central details as those that contribute to proving guilt of the suspect in the crime, while peripheral details were defined as being tangential to this but supporting witness credibility. She demonstrated that, across participants, a higher proportion of the central details was remembered than the proportion of the peripheral details. Interestingly, Shapiro also explored the effect of questioning format and found this pattern held regardless of whether questions were formulated as multiple choice or an open-ended response format. Multiple choice questions did, however, elicit a higher quantity of details compared with open-ended questions, but the response categories were limited to only one incorrect alternative and a "not sure" response (three response options in total).

The relationship between detail salience (central or peripheral) and repeated questioning has also been explored. Migueles and Garcia-Bajos (1999) compared students' eyewitness recall of a video robbery sequence using two free recall tests. They defined central information as that contained in the scenes depicting the criminal

event itself—a kidnapping—with content of the remaining video scenes defined as peripheral. They found that recall tended to improve with repeated questioning, but this was predominantly due to an increase in peripheral recall at time 2, with recall for central details improving by only a small degree. They note, "the first recall revolved around actions and in the second trial subjects repeated the narration, filling it in mainly with peripheral details" (Migueles and Garcia-Bajos 1999, p. 264). Migueles and Garcia Bajos further followed their free recall trials with a recognition test using statements that participants scored as true or false. Contrary to Shapiro's results above, participants showed greater accuracy for peripheral than central details; however, this was driven by the number of false positives. In other words, when provided with a plausible sentence and only two response options, participants tended to accept it as "true" rather than answer "false". The authors discussed the results in terms of the participants being misled by the false statements, particularly due to their "typicality" in relation to a normative script.

In relation to officer-involved shootings, the study by Alpert et al. (2012) explored police officers' memory for "threat-related" details and environmental details separately, finding officers recalled more threat than environmental details. There were also slight differences in the effect of timing of questioning across these two types of detail. This disaggregation of recall provides a useful first exploration of memory variation in the officer context, and the definition of threat adopted by Alpert et al. is consistent with the definition of central details often presented in eyewitness memory research; that is, descriptions of objects or people who may present a threat (details central to the offender and the offence). However, just as Shapiro (2006) defined central details within the role of the eyewitness (details that legally demonstrate the commission of the crime by the suspect), the definition of threat for police officers in an OIS needs to be considered alongside their role in the incident. Particularly, police officers must respond to the offender and make a decision as to whether to use lethal force or not. Police policy and training dictates conditions under which an officer's use of lethal force is reasonable and legally justified; typically, officers must specifically perceive imminent threat to human life. Our study seeks to adopt a police policy-relevant distinction between details that are threat-related (specifically pieces of information that affect the officers' safety and well-being at the event) and details not directly related to perception of threat, although contextually relevant to a later investigation (e.g. description of the shooter).

Threatening stimuli, however, are not always visual, and studies of police officers who have been involved in shootings have highlighted the variety of perceptual distortions that can affect recall. For example, Klinger (2007) and Klinger and Brunson (2009) report that, while officers often experience "tunnel vision", there are also frequent cases of auditory distortions and inaccurate reports of auditory information, particularly the number of shots fired (see also Alpert 1987), as well as distortions regarding spatial judgements, such as where the suspect was positioned in relation to the officer (see also Lewinski et al. 2016). While that research advances our knowledge about perceptual distortions and memory gaps after an OIS, it does not address the comparative accuracy of officers' recall or issues of question timing. In contrast, most studies on memory relevant to the OIS context (see the review of Grady et al. 2016) explore recall of visual stimuli rather than other sensory forms of detail such as auditory or spatial. Further, the study by Alpert et al., while assessing the difference between threat and non-threat details, did not ensure that comparisons controlled for the sensory

type of details within each category. Hartman et al. (2017), in their study of officers' memory for a training scenario, concluded that recall not only differs by type of detail but that recalling some types of detail (such as verbal interactions) was detrimental to recalling other "critical" information. Our study seeks to disaggregate recall by sensory type in order to more fully explore the effects of question timing on officers' recall.

The purpose of this study is, therefore, to add empirical findings to the debate over when to interview officers involved in an OIS (or other stressful event), and provide greater understanding of the limits of officers' memory. We explore three research questions relevant to the practical concern of the effects of question timing on accuracy of officers' memory: whether delaying questioning affects the accuracy of officers' memory for shooting events, whether officers' memory accuracy changes with repeated questioning and whether prior questioning improves later memory accuracy compared to no prior questioning. In answering each research question, our study disaggregates memory by the type of detail (by salience and sensory type), to explore whether some forms of detail are more accurately remembered than others and the interaction of this with questioning conditions. We also explore the role of emotional and cognitive state to add context to the officer experience and help understand any observed effects on memory.

Method

Sample

A total of 89 active sworn police officers took part in the study. However, the data for two individuals were removed due to a temporary break down in the random group assignment for these two officers (described later). Table 1 shows the demographic information for the 87 remaining participants. One participant did not disclose any demographic information and one further participant only disclosed his/her rank. The mean age of the sample was 42.32 years old with an average of 15.68 years of experience as a police officer. The majority of the sample was male (95.30%), white (94.10%), with some education post high school (65.90%), at the rank of senior constable (52.30%) or above (32.50%) with a usual role of general duties (i.e. patrol) officer (54.10%). As noted later in the procedure section, all participants were undergoing training to be police training instructors; thus, the sample is somewhat older and more experienced than the average officer. However, the ranks represented are still those involved in operational duties, and so most likely to encounter critical incidents such as the use of (lethal) force.

Design

The purpose of the randomised controlled trial was to test the effect of question timing on memory of information regarding involvement in an officer-involved shooting, as measured through participation in a simulated active armed offender (AAO) incident. Officers who are involved in a real shooting would typically be given an opportunity to provide a narrative "free recall" of the event, which would then be followed up with specific lines of questioning to ascertain particular details. We therefore included both free recall and follow-up questions in our design. The number of "details reported" through free recall was measured as a

Table 1 Sample by group

Variable	Total n = 87	Group 1 (immediate and repeated) n = 46	Group 2 (delayed) n = 41		
	Mean (SD)	Mean (SD)	Mean (SD)	t (df)	p
Age	42.32 (8.47)	42.36 (8.15)	42.28 (8.94)	0.04 (82)	0.969
Experience (years)	15.68 (8.84)	15.19 (9.22)	16.23 (8.47)	− 0.54 (84)	0.588
	Frequency	Frequency	Frequency	χ^2 (df)	p
Gender					
Male	81	44	37	0.03 (1)	0.627
Female	4	2	2		
Race					
White	80	42	38	0.06 (1)	0.565
Other	5	4	1		
Education					
High school	29	18	11	1.11 (2)	0.404
Some college	35	19	16		
University	21	9	12		
Rank					
Constable	11	6	5	0.03 (4)	1.000
Snr constable	45	24	21		
Sergeant	26	14	12		
Snr sergeant	2	1	1		
Other	2	1	1		
Duties					
General duties	46	24	22	0.45 (4)	0.978
Traffic	6	3	3		
Tactical crime squad	6	3	3		
Education and training officer	12	7	5		
Other	15	9	6		

dependent variable (DV). However, the primary dependent variable of interest for all three research questions was accuracy of memory. Due to potential issues in using free recall data to explore accuracy[1] for the purposes of our experimental manipulation, our accuracy measure was based on specific follow-up questions, with structured responses that were consistent across all participants. The primary DV, therefore, is correct recognition (or identification) of 19 details of the live action training scenario. This allows each to be

---

[1] The focus on details reported, rather than accuracy of the free recall, was chosen because assessments of accuracy of free recall can be affected by a number of individual factors (such as whether the participants thought a detail was relevant, how detailed they thought the description should be, fatigue, or individual language used in their reports, etc.). Such factors make between-subject comparisons of accuracy using these data difficult. As an example, one officer might report they fired 'five times', another 'between three and six times', while another may report they fired 'a string of shots'. Each of these responses may be correct, but each also demonstrates a different level of accuracy.

initially scored as a binary *incorrect* or *correct* response and then converted to a percentage score to explore the degree of accuracy.

Participants were randomly assigned (using SPSS random number generation) to one of two groups. Table 1 displays the characteristics of the sample in each group. There were no significant differences between the two groups on any of the demographics measured. The timing of questioning concerning an AAO scenario was manipulated for the groups, shown in Table 2. Officers assigned to group 1 were tested immediately after experiencing the field training scenario (condition 1), and again 2 days later (condition 3). Group 2 was tested only 2 days after the scenario (condition 2).[2] Thus, while the design is based on a 2 (group) × 2 (time) between-within subjects design, testing only occurred in three conditions of interest. Comparisons among pairs of these conditions enable the exploration of the three research questions, representing two between-subjects comparisons and one within-subjects comparison. Therefore, while analysed separately, they are not completely independent, as described below.[3]

The first factor in the design for each research question is the timing condition under which participants were tested. Research question 1 explores the effect of delaying questioning on memory for the shooting event. Timing condition is a two-level between-subjects factor that represents the comparison between the immediate recall of group 1 (condition 1) and the delayed recall of group 2 (condition 2). Research question 2 explores the effect of repeated questioning on officer memory (asking whether memory of the shooting event changes with repeated questioning). Timing condition is a two-level within-subjects factor representing the comparison between the immediate (condition 1) and repeated (condition 3) questioning of group 1. Research question 3 explores the effect of prior questioning on officer memory. Timing condition is a two-level between-subjects factor that represents the comparison between group 1 participants' recall at time 2, having been questioned prior (condition 3) and the recall of group 2 at time 2, not having been questioned prior (condition 2).

To assess the main effects of the treatment conditions (timing of questioning) on the DV (recognition memory accuracy), it was important to ensure that the content of the questions was also considered. The 19 questions concerned details that varied both as to the threat level of the information (two levels: threat or non-threat) and the sensory type of detail (three levels: visual, auditory or spatial/temporal). Generating questions under this factorial approach, therefore, ensured each question concerned either a threat or non-threat detail that was either visual, auditory or spatial/temporal in nature. All participants were questioned on all 19 details; therefore, *threat* and *sensory type* are

---

[2] Cook and Campbell (1979) define this post-test only randomised design as a variation of the classical experiment. It is preferable to the pre-post–experimental design when the pre-test has a significant likelihood of altering subjects' behaviour under observation (affecting measures at T2), or when tests conducted at T2 are contingent on knowledge gained while under observation (e.g. performing tasks, role-play or training scenarios).

[3] It is recognised that sensitive designs have a power level greater than 0.80, and at minimum, it is advised that statistical tests have a power level above 0.50. This indicates that a significant finding is more likely than not if the null hypothesis is false (see Weisburd and Britt 2007). Power analysis for this study reveals a power level of 0.94 for large effects (i.e. 38% difference), 0.61 for medium effects (i.e. 25% difference) and 0.15 for small effects (i.e. 10% difference). While the design is sensitive enough to detect statistical significance for large differences between the treatment and control groups, some caution should be used when considering moderate effects. Although statistical power is above the minimal level, it is below the optimal level used to define a statistically powerful study.

Table 2 Study design

Group	Event	Time 1: immediate	Time 2: 2 days later
Group 1	AAO scenario	Memory test (condition 1)	Memory test (condition 3)
Group 2	AAO scenario	No testing	Memory test (condition 2)

explored as within-subjects factors for all three research questions. The study design for each research question is, therefore, a 2 (timing condition) × 2 (threat) × 3 (sensory type) design, which allows for the main effect of the treatment conditions to be explored while controlling for the threat and sensory type of the information within the questions, as well as exploring interactions between these.

In addition, the study measured subjective anxiety and both subjective and objective cognitive ability. These measures are included to both check the effects of the experimental manipulation and as potential explanatory variables for any main effects observed through the above design. Specifically, we are interested to see if participants feel anxious after the scenario and whether this dissipates over time. Conversely, we are interested in whether cognition is impaired following the scenario and improves over time.

Materials

A written questionnaire was developed from previous research (see Alpert et al. 2012) to test accuracy of memory and included demographic questions. Cognitive tests were administered by two researchers. An observation checklist was also utilised to record participants' behaviour during the scenario.

Memory questionnaire

Free recall The survey first asked participants to describe what they could remember about the events that occurred in the scenario. Participants were directed to report details of events both leading up to the encounter with the shooter and during the encounter (i.e. report what they remember from before they were in the room with the shooter as well as while they were in the room with the shooter). This was to encourage the greatest amount of narrative detail and to resemble writing of an incident report.

Recognition Memory accuracy was measured through 19 structured response questions; therefore, the specific mechanism tested here is recognition memory. Each question related to a different detail of the event, which varied according to the threat level as well as the sensory type of information. The threat level of the information describes the importance, or salience, of the detail and was categorised into two mutually exclusive categories of threat and non-threat details. Threat details were defined as details that would be relevant to the immediate safety and physical well-being of the officers (or victims) in responding to the incident. Non-threat items related to details that fell outside of this definition, but were contextual details relevant to an investigation of the incident. Items were generated through a factorial approach within the 2 (threat) × 3 (sensory type) design, which produced six categories of questions. Initial generation of questions was based on consultation with three

subject matter experts (police training officers). Questions were then categorised independently by the three authors to ensure agreement and reliability.

Nine items were agreed to concern threat details: three visual items, incorporating details about weapons present and victim injury; three auditory items concerning verbal threats to life; and three spatial items concerning the proximity of the shooter to the officers. Ten items were agreed to be non-threatening details (not immediate safety concerns, but of interest to investigators): three visual items included clothing colours of those present and an object the victim was holding; four auditory items concerned which officer fired first, the number of shots the officers each fired at the shooter and victim speech; and three spatial/temporal items asked how long the scenario lasted, the number of rooms the officers searched before they encountered the shooter and where the shooter was standing.

Cognitive tests

Objective measures of memory and cognition Following the procedure of the study of White et al. (2015) on Taser exposure, participants were administered the Hopkins Verbal Learning Test (HVLT) (Brandt and Benedict 2001), a Trail Making test (see Tombaugh 2004) and the Weschler (1945) Digit Span test. On all tests used, higher scores indicate higher cognitive ability. The HVLT tested verbal learning and memory through recall of a list of 12 words over three consecutive trials, followed by a delayed fourth trial (after approximately 10 min). The number of correctly recalled words on trials 1 to 3 was summed for a *total recall* score. The *percent retained* at trial 4 was calculated as the number of correct words at trial 4 divided by the number of correct words at trial 3, multiplied by 100. Finally, participants were asked to recognise the 12 words from a list of 24 words. A *discrimination index* was calculated as the number of true positives minus the number of false positives. Two versions of the HVLT test have been validated, each using a different set of words. All participants received version 1. Additionally, version 2 was used with participants who were tested a second time. This meant that within-subject comparisons—between the first and second testing—could be made while reducing the effects of long-term memory/learning from having completed the same test on another occasion.

The second cognitive test timed participants on two trail making tasks that measure speed of processing and executive functioning. Participants must draw a line between consecutive numbers (or alternating letters and numbers) as quickly as possible. Times were summed for each participant for a total time. Finally, the digit span test measured short-term auditory learning by requiring participants to repeat progressively longer strings of numbers forwards and then backwards. Scores were summed for a total of between 0 (no strings of numbers recited correctly) and 30 (all 30 number strings recited correctly).

Subjective measures of memory and concentration Participants were asked to rate their general ability to remember things (now and on a normal day) and their ability to concentrate (now and on a normal day). The 10-point response scales for these items were anchored at 1 = *poor* and 10 = *very good*. Indexes of memory difference (from baseline) and concentration difference were calculated by subtracting the rating for *on a normal day* from the rating for *now*. A positive difference score therefore indicates

a better current ability compared to normal, while a negative score indicates they feel they are worse than normal.

Anxiety Deffenbacher et al. (2004) and Deffenbacher (1994) describe the stress response typically experienced by eyewitnesses to a crime as the result of the dominance of "activation mode" of attention control. Deffenbacher et al. (2004, pp. 287–288) state: "Tasks eliciting activation mode dominance include any task serving to increase cognitive anxiety (worry) and/or somatic anxiety (conscious perception of physiological activation), including vigilance, escape, avoidance, or "pressure" tasks". It is assumed that participants will experience heightened stress while they are undertaking the scenario (see below). However, this study is concerned with participants' state of mind at the time of questioning. Participants were therefore asked by the researchers two questions to measure their current and typical level of anxiety: "How anxious do you feel right now?" and "How anxious do you feel on a normal day?", both measured on a 10-point response scale with anchor points for 1 = *no anxiety* and 10 = *a great deal of anxiety*. An index of *anxiety differrence* (from baseline) was created by subtracting the latter from the former. A positive score on the index indicates participants felt more anxious than usual.

Observation checklist

An observation checklist was used to record the behaviour of officers as they undertook the scenario. The checklist consisted of a one-page form with duplicated items (and their response options) from the survey that could not be anticipated or controlled by the researchers (for example, the number of shots fired, or number of rooms entered). These details were recorded and used later to score the relevant questions for each officer.

Procedure

Participants were undergoing police service AAO training at the time of the research, which incorporated their responding to a live action role-play scenario involving a shooter and a victim in an abandoned building. The course was designed to train the officers to become AAO training instructors.[4] Before officers began the scenario-based portion of the training, the researchers obtained their written informed consent to participate in the study. While officers were obliged to undertake the training, participation in the study (testing of their memory) was voluntary. All officers gave consent to participate.

Officers responded to the scenario in pairs and were tasked with entering the building, locating an offender (and victim) in one of the upstairs rooms and resolving the incident by identifying and removing the threat (typically shooting the armed offender[5]). The scenario,

---

[4] The instructor course includes the material that all trainees (non-instructors) would be exposed to, plus additional material at the end of the program exclusively for those training to be instructors (received after contact with this research study).

[5] Officers were free to choose whether to shoot the offender or not. The majority of the officers fired at least one shot at the offender, with only eight out of the 87 (four in group 1 and four in group 2) not firing their weapon. However, on these eight occasions, the officer's partner fired. Thus, all scenarios ended with at least one of the officers discharging their weapon (resulting in an OIS).

therefore, likely invoked both physical exertion and cognitive anxiety as officers moved towards a source of threat, under pressure to resolve the incident, without knowing what they would encounter as they searched the building. Scenario-based training has been shown to simulate the psychological stress conditions of use of force events, even if force is only anticipated but not used (Armstrong at al. 2014). An observer documented key details throughout the scenario with the observer checklist and recorded the officers' actions and timed the scenario.

At the conclusion of the scenario, officers assigned[6] to group 1 were brought to a private interviewing room where the first and second authors administered the cognitive tests and memory questionnaire (condition 1, immediate recall), while those in group 2 continued their training program as usual. All participants returned 2 days later to be tested by the researchers[7]; this was the only time group 2 participants were tested (condition 2, delayed recall), while those in group 1 were tested for a second time (condition 3, repeated recall).[8] Testing of both groups, and at both times, followed the same procedure; thus, only the timing differed across the groups and conditions. Officers were tested one-on-one by either the first or second author using the materials described above. Researchers verbally asked the participants questions regarding their current emotional and cognitive state and administered the three cognitive tests,[9] taking between 10 and 15 min. The Trail Making and Digit Span tests were administered between the HVLT third and fourth (delayed) recall trial. Participants then completed the written memory survey, which first comprised a free recall section asking participants to write down what they remembered from the scenario, followed by the structured response recognition items and demographic questions.

Responses to the recognition survey items were scored as incorrect or correct. The observation checklist was used to ascertain the correct answers for details that were dependent upon the behaviour of the participants, and so could not be controlled by the researchers (such as the numbers of shots fired by participants, the number of rooms entered and how long the scenario took to complete). Data were entered into SPSS. All data were entered independently by two coders (i.e. double data entry); inconsistencies were checked and rectified to produce a final SPSS data file.

Analysis

Manipulation checks

As noted earlier, the position that delay in questioning may enhance recall is partly based upon the role of sleep in memory consolidation and partly on the notion that time allows a decrease in stress. While it is not the purpose of this paper to test these specific

---

[6] Random allocation was done in advance and applied on the day.

[7] One participant from group 1 was unable to return to be questioned a second time, leaving $n = 45$ repeated measures across conditions 1 and 3.

[8] Officers were asked not to discuss their actions in the scenario with one another before or between testing. However, there were no procedures in place to ensure these instructions were followed, such as isolating officers for the full 2 days.

[9] As noted in the "Method" section, participants in group 1 who were tested twice were given a different version of the HVLT at time 2 compared to time 1 to reduce learning effects. The trail making and digit span tests were the same at all times.

processes, it is useful to know whether our manipulated conditions that are separated by two sleep cycles also represent a significant difference in experienced emotional and cognitive state. To this end, independent and paired t tests were used to conduct manipulation checks to test whether the conditions differed by reported anxiety and subjective and objective measures of cognition.[10]

Analysis of free recall narratives

The free recall narratives collected in condition 1 (immediate) and condition 2 (delayed) were coded[11] to measure the number of details respondents mentioned. Rather than conducting a narrowly focused content analysis, coding the narratives required consideration of the broader contextual backdrop that shaped specific pieces of text, as well as the broader police culture and technical vernacular that impacts how officers describe such an event. Coding was conducted in two rounds. The first round was conducted to search for, and code, information representing the 19 details subject to later questioning (for example, whether the narrative provided any reflection on the number of shots fired by the officer). Thus, each narrative was coded for how many of the nine prescribed threat, and 10 prescribed non-threat, details were mentioned (each calculated as a proportion). Second, the presence of additional details (beyond those 19) was also coded. Each new detail mentioned was determined to be either threat relevant or non-threat relevant, in line with the definition provided earlier. The total number of threat and non-threat details was then calculated for each narrative. T tests were used to conduct initial comparisons between condition 1 and condition 2 regarding the total number of threat and total number of non-threat details mentioned, as well as the proportion of threat and non-threat details that were the focus of later questioning (the latter provides more direct comparability between the qualitative report data and the quantitative accuracy data). This exploration of the free recall data was then followed up with analysis of the structured response questions to explore the three research questions in relation to the DV of memory accuracy, described below.

Analysis of recognition memory accuracy

The structured response questions were scored and the proportion of correct answers that were recalled overall, and by category of information (threat × sensory type), was calculated as a percentage for each participant in each condition. Before conducting the analysis for the three research questions, the dependent variable of *mean percentage correct* was checked for dependencies between participant pairs (as noted in the procedure, all participants experienced the scenario in teams of two). Following Hope et al. (2016), the method of Alferes and Kenny (2009) was followed, using the syntax

---

[10] A paired t test was used to test the within-subjects difference in cognitive ability using only the HVLT. As noted in the "Method", the two versions of the HVLT were used in the different within-subjects conditions to reduce learning effects. Such alternate versions were not available for the other two cognitive tests and so within-subjects (over time) comparisons are not made for these tests as differences would be affected by learning.

[11] The narratives collected in condition 3 (repeated) were not analysed here because they may have been affected by the earlier specific questioning, and the impact of specific questioning on later free recall was not the focus of this study.

provided in their appendix (and supplementary materials) for indistinguishable pairs. Tests indicated the presence of dependency between paired responses. We therefore used linear mixed models (using SPSS MIXED with maximum likelihood estimation) to allow individuals' responses to be nested by "team" (pair). For each of the three research questions, the 2 (condition) ×2 (threat) × 3 (sensory) factorial design was analysed, entering all independent variables (main effects) and two-way and three-way interactions as fixed effects, with random effects specified for the repeated responses of the participants nested within the participant pairs. Likelihood ratio tests indicated that the nested models were a better fit to the data than the non-nested models (with team accounting for 6.74, 9.92 and 6.94% of the variance in recall for the three models, respectively), although the actual pattern of significance across the model effects (and, therefore, the conclusions drawn) was unchanged between the nested and non-nested models. Significant effects were followed up with pairwise comparisons of the estimated marginal means with Bonferroni correction, within the SPSS MIXED procedure, thus taking into account the model effects, rather than performing t tests on the raw group means. Only the results of interactions relevant to the research questions are reported in the Results section; interactions that do not include *condition* (i.e. threat x sensory) are reported in the Appendix.

Results

The study sought to explore the effect of timing of questioning on memory. First, this section explores the participants' present emotional and cognitive state in the conditions. Next, findings from the free recall data are presented. Finally, each of the three research questions is explored in turn: (1) the effect of delay, (2) the effect of repeated questioning and (3) the effect of prior questioning on memory accuracy.

Manipulation checks: self-reported emotional state and cognitive ability

Participants in each condition were compared regarding their subjective emotional state, using self-reported difference (to how they feel *on a normal day*) in anxiety, memory and concentration, as well as objective scores on cognitive tests (Table 3). At the time of questioning, participants in the immediate recall condition (condition 1) reported feeling significantly more anxious than usual and less able to remember and to concentrate than usual, compared to those in the delayed recall condition (condition 2). However, these differences should be treated with caution due to the number of multiple comparisons made. Group 1 participants also felt more anxious than normal immediately after the event (condition 1) compared to how they felt 2 days later (condition 3). Similarly, they also felt their ability to remember was worse than *on a normal day*, more so immediately compared to 2 days later. These differences remain significant even after applying Bonferroni-corrected p values. There was no significant within-subjects difference in the measure of concentration over time. There is some evidence, then, that the scenario adversely affects officers' subsequent subjective ratings of their emotional state and cognitive ability and that the effects dissipate over time.

Additionally, we explored whether participants' cognition differed across the conditions as measured by a series of objective tests. There were no significant differences on any of the

Table 3 Measures of anxiety, memory and cognition for the sample and comparison by condition

Measure	Condition 1 (group 1 immediate) n = 46 Mean (SD)	Condition 2 (group 2 delayed) n = 41 Mean (SD)	Condition 3 (group 1 r epeated) n = 45 Mean (SD)	Independent t test (C1 × C2)		Paired t test (C1 × C3)	
				t (df)	p	t (df)	p
Anxiety difference	1.22 (2.14)	0.27 (1.78)	0.20 (1.66)	2.23 (85)	0.028	2.77 (44)	0.008
Memory difference	−0.93 (.95)	−0.39 (.86)	−0.36 (0.88)	−2.78 (85)	0.007	−3.01 (44)	0.004
Concentration difference	−0.87 (1.76)	−0.15 (1.26)	−0.60 (1.29)	−2.22 (81.29)	0.029	−0.82 (44)	0.417
HVLT							
Total recall	24.35 (3.00)	25.66 (3.93)	24.42 (3.62)	−1.76 (85)	0.082	0.00 (44)	1.000
% retained	94.13 (15.56)	96.25 (25.07)	94.36 (17.09)	−0.48 (85)	0.633	0.04 (44)	0.965
Discrimination	10.41 (1.20)	10.32 (1.99)	10.22 (1.18)	0.28 (85)	0.784	0.77 (44)	0.445
Trail making	73.20 (19.57)	80.71 (31.31)	–	−1.36 (85)	0.178	–	–
Digit span	16.83 (3.32)	18.00 (3.42)	–	−1.62 (85)	0.108	–	–

Anxiety difference = the difference between self-reported ratings of anxiety 'now' minus 'on a normal day'; *memory difference* = the difference between self-reported ratings of memory ability 'now' minus 'on a normal day'; *concentration difference* = the difference between self-report ratings of ability to concentrate 'now' minus 'on a normal day'; *HVLT* = Hopkins Verbal Learning Test

cognitive tests between the immediate and delayed condition (condition 1 versus condition 2). The repeated measures for officers in group 1 on the HVLT also did not differ (comparing when tested immediately (condition 1) and 2 days later[12] (condition 3)). Thus, officers' general cognition (their ability to respond on the tests) did not seem to be directly affected by how recently they had experienced the scenario, and no significant improvement was evident after 2 days either between or within groups.

Does timing affect the amount of details mentioned in free recall narratives?

The amount of details reported in the free recall narratives was compared between condition 1 (immediate report) and condition 2 (delayed report). The mean total number of threat details reported in the narratives did not differ significantly across the conditions (condition 1 mean = 5.02 (SD = 1.36); condition 2 mean = 4.44 (SD = 1.75); $t(85) = 1.75$, $p = 0.084$). This was also true when looking only at the proportion of the nine details that were later subject to specific questioning (condition 1 mean = 35.51% (SD = 11.38); condition 2 mean = 31.44% (SD = 13.81); $t(85) = 1.51$, $p = 0.136$). In contrast, the reports given immediately mentioned significantly more non-threat details than those given 2 days later, both for total number of non-threat details mentioned (condition 1 mean = 9.10 (SD = 2.50); condition 2 mean = 7.00 (SD = 2.67); $t(85) = 3.724$, $p < 0.001$), and for the proportion of the 10 non-threat details about

---

[12] As noted earlier, the HVLT at time 2 was a different version to time 1. The trail making and digit span tests are not tested for within-group differences since the same versions were used at both times and improvement would be expected due to leaning effects.

which they were subsequently questioned (condition 1 mean = 33.70% (SD = 11.99); condition 2 mean = 24.63% (SD = 15.02); $t(85) = 3.125$, $p = 0.002$).

Does timing of questioning affect memory accuracy?

To answer research question 1, we used a linear mixed model to test the main, and interaction, effects of timing condition (condition 1: immediate, versus condition 2: delayed questioning) and question content regarding threat (threat, non-threat) and sensory type (visual, auditory and spatial/temporal) on memory accuracy for the scenario. Table 4 shows the estimated marginal means. The main effect for timing was not significant, $F(1, 45.61) = 2.13$, $p = 0.151$. The main effect for threat type was significant $F(1, 478.34) = 6.86$, $p = 0.009$, with participants tending to remember a significantly higher proportion of threat items than non-threat items. There was also a significant main effect of sensory type, $F(2, 478.34) = 49.86$, $p < 0.001$. Pairwise comparisons showed participants remembered a significantly higher proportion of visual details compared to auditory details, mean difference = 21.98, SE = 2.64, $t(1, 478.34) = 8.34$ $p < 0.001$, 95% CI [15.64, 28.31], as well as spatial/temporal details, mean difference = 23.54, SE = 2.64, $t(1, 478.34) = 8.93$, $p < 0.001$, 95% CI [17.21, 29.88].

There was a significant two-way interaction between condition and threat type, $F(1, 478.34) = 8.90$, $p = 0.003$ (see Fig. 1). Pairwise comparisons revealed that there was no significant difference between the immediate and delayed conditions for recall of threat-related information, mean difference = $- 2.15$, SE = 3.62, $t(1, 105.14) = -0.59$, $p = 0.556$, 95% CI [$- 9.35$, 5.06]. However, recall of non-threat-related information was significantly better in the immediate condition compared with the delayed condition, mean difference = 10.70, SE = 3.62, $t(1, 105.14) = 2.94$, $p = 0.004$, 95% CI [3.49, 17.90]. The interaction between condition and sensory type was not significant, $F(2, 478.34) = 2.33$, $p = 0.098$. While not hypothesised, there was a significant interaction of threat with sensory type (see Appendix).

In answer to research question 1, then, timing of questioning did not impact memory accuracy for types of details equally; the effect was moderated by the threat content of the questioned material. The group questioned 2 days after the incident (condition 2) had poorer recognition of non-threat-related details than the group questioned immediately, but delay did not impair recognition of threat-related information.

Repeated questioning: does delay change memory?

This section explores within-subjects' change in memory to address research question 2; asking whether the individuals in group 1 who were questioned immediately after the incident (condition 1) showed a difference in their memory when questioned 2 days later (condition 3). A linear mixed model explored the condition × threat × sensory type factorial design, where condition represents the repeated memory test of group 1. Table 5 shows the estimated marginal means of the percentage of questions answered correctly.

There was no significant main effect of timing ($F(1, 505.12) = 0.22$, $p = 0.643$). Thus, participants' accuracy level did not significantly change over the 2 days. There was also no significant main effect of threat content of the question ($F(1, 500.44) = 0.03$, $p = 0.861$) on recall; those in group 1 did not remember threat details significantly more than non-threat details over all. There was a significant main effect of sensory type of detail on accuracy ($F(2, 500.44) = 94.35$, $p < 0.001$). Pairwise comparisons

Table 4 Estimated marginal means, standard errors and 95% confidence intervals for research question 1: condition (timing) × threat × sensory type

Condition	Threat level	Visual		Auditory		Spatial/temporal		Total	
		Mean (SE)	95% CI	Mean (SE)	95% CI	Mean (SE)	95% CI	Mean (SE)	95% CI
Immediate (cond. 1)	Threat	77.58 (3.87)	[69.98, 85.18]	23.95 (3.87)	[16.35, 31.56]	44.24 (3.87)	[36.64, 51.85]	48.59 (2.49)	[43.65, 53.53]
	Non-threat	54.39 (3.87)	[46.79, 61.99]	53.85 (3.87)	[46.24, 61.45]	39.90 (3.87)	[32.29, 47.50]	49.38 (2.49)	[44.44, 54.32]
	Total	65.98 (2.90)	[60.27, 71.70]	38.90 (2.90)	[33.18, 44.62]	42.07 (2.90)	[36.35, 47.79]	**48.98 (2.01)**	**[44.95, 53.02]**
Delayed (cond. 2)	Threat	74.72 (4.10)	[66.67, 82.78]	36.51 (4.10)	[28.45, 44.57]	40.98 (4.10)	[32.93, 49.04]	50.74 (2.64)	[45.50, 55.98]
	Non-threat	41.39 (4.10)	[33.33, 49.45]	45.86 (4.10)	[37.80, 53.92]	28.79 (4.10)	[20.73, 36.85]	38.68 (2.64)	[33.44, 43.92]
	Total	58.06 (3.07)	[51.99, 64.12]	41.19 (3.07)	[35.12, 47.25]	34.89 (3.07)	[28.82, 40.95]	**44.71 (2.13)**	**[40.42, 49.00]**
Combined	Threat	76.15 (2.82)	[70.61, 81.69]	30.23 (2.82)	[24.69, 35.77]	42.61 (2.82)	[37.08, 48.15]	**49.67 (1.82)**	**[46.06, 53.27]**
	Non-threat	47.89 (2.82)	[42.35, 53.43]	49.85 (2.82)	[44.31, 55.39]	34.34 (2.82)	[28.80, 39.88]	**44.03 (1.82)**	**[40.43, 47.63]**
	Total	**62.02 (2.11)**	**[57.85, 66.19]**	**40.04 (2.11)**	**[35.88, 44.21]**	**38.48 (2.11)**	**[34.31, 42.64]**	*46.85 (1.46)*	*[43.90, 49.79]*

Bold indicates means for main effects; italics indicate means for two-way interactions; bold italics indicate the grand mean; remaining (unformatted) means are for the three-way interaction
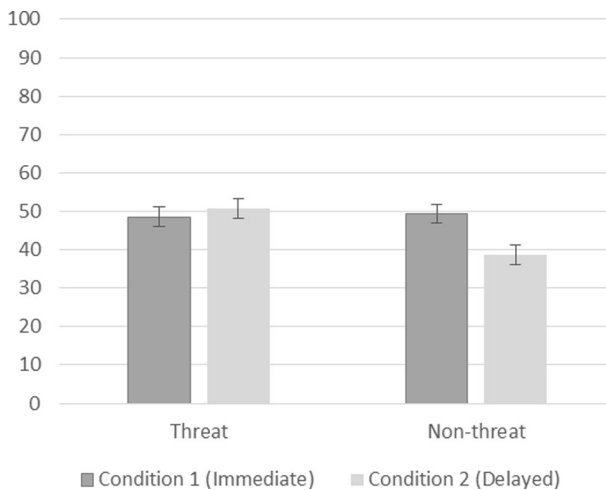
Fig. 1 Mean percentage of correct answers and standard errors, for the two-way interaction of threat by timing (condition 1 versus condition 2)

showed that visual details were better recognised than both auditory (mean difference = 28.26, SE = 2.32, $t(1, 500.44) = 12.18$, $p < 0.001$, 95% CI [22.70, 33.83]) and spatial/temporal (mean difference = 26.77, SE = 2.32, $t(1, 500.44) = 11.54$, $p < 0.001$, 95% CI [21.21, 32.33]) details. The hypothesised interactions for condition were not significant: condition with threat, $F(1, 500.44) = 0.06$, $p = 0.810$; condition with sensory type, $F(2, 500.44) = 0.77$, $p = 0.464$; or the three-way interaction between condition, threat and sensory type, $F(2, 500.44) = 0.07$, $p = 0.934$. Again, there was a significant interaction of threat with sensory type (see Appendix).

However, any change in the officers' answers over repeated questioning may be significant in a practical sense in an investigation, even if not statistically significant. All 19 questions had at least one participant who changed his/her answer between the two questioning times. Four questions showed a change of answer by at least 20% of the sample: number of rooms entered, number of weapons present, length of time, and colour of the shooter's shirt. Overall, from a total of 855 pairs of responses (45 participants × 19 items), 89% of the responses were consistent across the two questioning times, 5% of responses changed from a correct to an incorrect answer and 6% changed from an incorrect to a correct answer. This averages to 2.13 answer changes per participant.

In answer to research question 2, then, officers who were questioned immediately after the event showed similar levels of memory accuracy (at the aggregate level) when questioned again 2 days later. This suggests that questioning early helps mitigate the memory decay that we saw for non-threat details in the delayed condition in the previous section. This conclusion will be illuminated further in the next section. However, there were still some differences in the participants' responses across the two time points, with some participants changing their answers for the better and some for the worse.

Does prior questioning improve later memory accuracy?

Research question 3 concerns the effect of prior questioning on memory decay; that is, do people who have previously been questioned provide more accurate responses

Table 5 Estimated marginal means, standard errors and 95% confidence intervals for research question 2: condition (repeated questioning) × threat × sensory type on memory accuracy

Condition	Threat level	Visual		Auditory		Spatial/temporal		Total	
		Mean (SE)	95% CI	Mean (SE)	95% CI	Mean (SE)	95% CI	Mean (S.E.)	95% CI
Immediate (cond. 1)	Threat	77.62 (3.63)	[70.47, 84.78]	23.99 (3.63)	[16.85, 31.15]	44.29 (3.63)	[37.13, 51.44]	48.64 (2.48)	[43.70, 53.57]
	Non-threat	54.43 (3.63)	[47.28, 61.59]	53.89 (3.63)	[46.74, 61.04]	39.94 (3.63)	[32.79, 47.10]	49.42 (2.48)	[44.49, 54.35]
	Total	66.03 (2.81)	[60.46, 71.59]	38.94 (2.81)	[33.38, 44.51]	42.12 (2.81)	[36.55, 47.68]	**49.03 (2.09)**	**[44.80, 53.26]**
Repeated (cond. 3)	Threat	82.56 (3.67)	[75.34, 89.78]	25.52 (3.67)	[18.30, 32.74]	41.82 (3.67)	[34.60, 49.04]	50.00 (2.50)	[45.00, 54.94]
	Non-threat	56.64 (3.67)	[49.42, 63.86]	54.78 (3.67)	[47.56, 62.00]	38.12 (3.67)	[31.00, 45.34]	49.85 (2.50)	[44.88, 54.81]
	Total	69.60 (2.83)	[63.99, 75.21]	40.15 (2.83)	[34.55, 45.76]	39.97 (2.83)	[34.36, 45.58]	**49.91 (2.10)**	**[45.66, 54.16]**
Combined	Threat	80.09 (2.82)	[74.51, 85.68]	24.76 (2.82)	[19.18, 30.35]	43.05 (2.82)	[37.47, 48.64]	**49.30 (2.09)**	**[45.07, 53.54]**
	Non-threat	55.53 (2.82)	[49.95, 61.12]	54.33 (2.82)	[48.75, 59.92]	39.03 (2.82)	[33.44, 44.61]	**49.63 (2.09)**	**[45.40, 53.87]**
	Total	**67.81 (2.30)**	**[63.21, 72.42]**	**39.55 (2.30)**	**[34.95, 44.15]**	**41.04 (2.30)**	**[36.44, 45.64]**	*49.47 (1.87)*	*[45.62, 53.32]*

Bold indicates means for main effects; italics indicate means for two-way interactions; bold italics indicate the grand mean; remaining (unformatted) means are for the three-way interaction

2 days after the event than those who have not been previously questioned? Again a linear mixed model is used to test the 2 (timing condition) × 2 (threat) × 3 (sensory type) design. For this analysis, the two levels of timing condition are condition 2 (group 2 officers questioned after 2 days, for the first time) as compared to condition 3 (group 1 officers questioned after 2 days, for the second time). Table 6 shows the estimated marginal means of the percentage of correct answers.

The main effect of timing condition (repeated versus delayed) was not significant ($F(1, 43.33) = 2.99$, $p = 0.091$). The main effect of threat was significant ($F(1, 470.31) = 7.97$, $p = 0.005$), with threat details recalled more than non-threat details. The main effect of sensory type was also significant ($F(2, 470.31) = 59.33$, $p < 0.001$). Pairwise comparisons showed that visual details were better recalled than both auditory (mean difference = 23.16, SE = 2.64, $t(1, 470.31) = 8.77$, $p < 0.001$, 95% CI [16.81, 29.51]) and spatial/temporal information (mean difference = 26.40, SE = 2.64, $t(1, 470.31) = 10.00$, $p < 0.001$, 95% CI [20.05, 32.75]).

The two-way interaction of condition with threat was significant ($F(1, 470.31) = 7.65$, $p = 0.006$) (see Fig. 2). Pairwise comparisons revealed that condition (presence/absence of prior questioning) did not affect recognition of threat-relevant details (mean difference = − 0.87, SE = 3.66, $t(1, 99.73) = − 0.24$, $p = 0.813$, 95% CI [− 8.12, 6.39]), but those who had been questioned previously had better recognition of the non-threat items (mean difference = 11.07, SE = 3.66, $t(1, 99.73) = 3.02$, $p = 0.003$, 95% CI [3.82, 18.32]). The interaction of condition with sensory type was not significant ($F(2, 470.31) = 2.83$, $p = 0.060$), neither was the three-way interaction between condition, threat and sensory type ($F(2, 470.31) = 0.86$, $p = 0.425$). The interaction of threat with sensory type was again significant (see Appendix).

In answer to research question 3, then, officers who had been questioned immediately after the event showed better recognition of the non-threat items after 2 days than did those officers who had not been questioned earlier. Thus, early questioning improved memory retention for non-threat-related items.

Discussion

This study adds to the sparse applied empirical literature suggesting how long after a critical incident or OIS police officers should be interviewed to retrieve the most details from memory. Competing perspectives over interview timing argue between interviewing officers as soon as possible, to prevent memory decay and contamination or, conversely, giving officers time to rest and consolidate their memory to improve accuracy. Our study supports the former perspective, although the effect of delay on memory, and reporting, was dependent on the type of information subject to questioning. We did not find any evidence that delay improves memory, degree of reporting or cognitive capability that could indicate enhanced ability to respond to questioning.

It was not the purpose of this paper to directly test the underlying arguments for a delayed interview, such as the role of sleep in memory consolidation, or the direct effect of anxiety at retrieval on memory. However, while stress at the time of encoding affects memory for the incident (Hope 2016), our results suggest that the anxiety experienced at the time of questioning (retrieval) did not seem to be related to officers' ability to report, or correctly recognise, details of the scenario, or perform in the cognitive tasks. Participants felt

Table 6 Estimated marginal means, standard errors and 95% confidence intervals for research question 2: condition (prior versus no prior questioning) × threat × sensory type

Condition	Threat level	Visual		Auditory		Spatial/temporal		Total	
		Mean (SE)	95% CI	Mean (SE)	95% CI	Mean (SE)	95% CI	Mean (SE)	95% CI
Repeated (cond. 3)	Threat	82.47 (3.90)	[74.79, 90.13]	25.43 (3.90)	[17.75, 33.10]	41.72 (3.90)	[34.05, 49.40]	49.87 (2.52)	[44.88, 54.87]
	Non-threat	56.54 (3.90)	[48.87, 64.21]	54.69 (3.90)	[47.01, 62.36]	38.02 (3.90)	[30.35, 45.69]	49.75 (2.52)	[44.75, 54.74]
	Total	*69.50 (2.93)*	*[63.73, 75.28]*	*40.06 (2.93)*	*[34.28, 45.83]*	*39.87 (2.93)*	*[34.10, 45.65]*	**49.81 (2.03)**	**[45.72, 53.90]**
Delayed (cond. 2)	Threat	74.72 (4.10)	[66.67, 82.78]	36.51 (4.10)	[28.46, 44.57]	41.39 (4.10)	[32.93, 49.04]	50.74 (2.65)	[45.48, 56.00]
	Non-threat	41.39 (4.10)	[33.33, 49.44]	45.86 (4.10)	[37.81, 53.91]	28.79 (4.10)	[20.73, 36.84]	38.68 (2.65)	[33.42, 43.94]
	Total	*58.06 (3.08)*	*[51.98, 64.13]*	*41.19 (3.08)*	*[35.11, 47.26]*	*34.88 (3.08)*	*[28.81, 40.96]*	**44.71 (2.14)**	**[40.39, 49.03]**
Combined	Threat	*78.59 (2.83)*	*[73.03, 84.16]*	*30.97 (2.83)*	*[25.41, 36.53]*	*41.35 (2.83)*	*[35.79, 46.92]*	**50.31 (1.83)**	**[46.68, 53.93]**
	Non-threat	*48.96 (2.83)*	*[43.40, 54.53]*	*50.27 (2.83)*	*[44.71, 55.84]*	*33.40 (2.83)*	*[27.84, 38.97]*	**44.21 (1.83)**	**[40.59, 47.84]**
	Total	**63.78 (2.12)**	**[59.59, 67.97]**	**40.62 (2.12)**	**[36.43, 44.81]**	**37.38 (2.12)**	**[33.19, 41.57]**	***47.26 (1.48)***	***[44.28, 50.23]***

Bold indicates means for main effects; italics indicate means for two-way interactions; bold italics indicate the grand mean; remaining (unformatted) means are for the three-way interaction
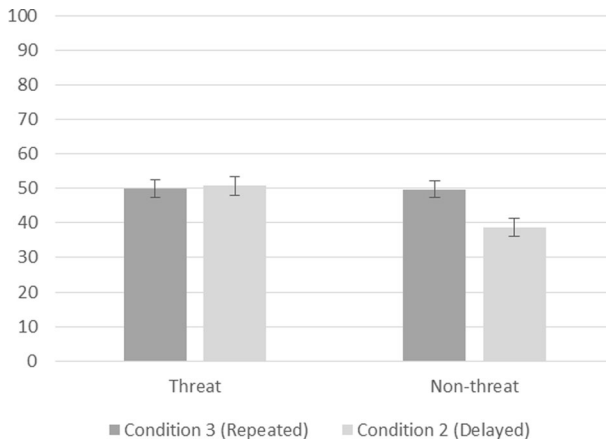
Fig. 2 Mean percentage of correct answers and standard errors, for the two-way interaction of threat by condition of no prior questioning (condition 2) versus prior questioning (condition 3)

heightened anxiety and reported less confidence in their cognitive ability, immediately after the high-stress active armed offender training compared with 2 days later. This indicates the immersive effects of the simulation, and that time improves subjective state. However, this pattern of improvement over time was not reflected in the objective cognitive tests, or in the analyses of memory accuracy; while the officers perceived the benefits of delay, their actual performance was either unchanged or worse. Our assessment of anxiety was limited, though, being based on self-report. Further, the officers had undergone a simulation rather than a real shooting event. While anxiety levels immediately after the scenario were statistically significantly higher than the officers' typical (and later) levels, the mean difference recorded was relatively small in absolute terms (given the possible range of the scale we used). Thus, while our results suggest immediate questioning is better for memory, this must be considered alongside other priorities, such as protecting and optimising officer well-being.

The non-significant main effect of timing on recognition suggested that, despite reductions in anxiety (and the experience of two sleep cycles), delay does not improve the accuracy of memory, in line with the conclusions of Grady et al. (2016). However, the presence of the interaction with the type of details being recalled revealed a more nuanced picture regarding decay of memory over time. Threat-relevant memory (recognition of details relevant to the immediate safety and well-being of the officers and victim) was not significantly affected by the timing of questioning. However, memory for non-threat information decayed over time, if officers were not prompted early to remember such details. This has implications for identifying the best time to interview officers. Policy makers may wish to consider timing of interviews in light of different interview objectives, rather than a "one size fits all" practice. One could argue that threat-relevant details are particularly pertinent to an investigation into an officer's decision to shoot, and questioning for these details could be delayed (for welfare or logistical reasons, etc.) without losing accuracy. However, our non-threat details were facts that investigators would want to establish to understand the broader context of the incident and thoroughly investigate the case, and these were subject to memory decay if not prompted with questions immediately. Prior questioning, however, mitigated this memory decay, in line with the testing effect (Roediger III and Karpicke 2006) and

research that suggests "priming" recall with early targeted questioning, or a non-leading interview, can improve later recall for that information (see review of Grady et al. 2016; Powell and Thomson 1997). Some US jurisdictions subscribe to a policy of conducting at least a brief interview, or "safety statement", soon after the event to establish some initial information regarding issues of public safety. Our results suggest that this might be beneficial for memory of non-threat details.

Nevertheless, some changes in officers' memory were apparent over time and so should be expected at the individual level. This finding has important implications for interpretations of officers' motives when they change their answers after being questioned repeatedly. In the context of a real shooting, inaccuracies in recall, and particularly changes in answers over time, can give rise to suspicion that officers are deliberately misleading investigators or presenting a self-serving version of events (Klinger and Brunson 2009). Our findings show that changes in answers are not uncommon, even under simulated conditions, and so can be considered a likely product of memory errors and, potentially, lack of confidence in memory. These changes in answers cannot necessarily be attributed to conscious deception (Hope 2016).

Further, regardless of timing or type of detail subject to questioning, officers on average recalled less than half the questions correctly, and for some categories of information under certain conditions, this dropped further. Thus, just as eyewitness memory can be impaired for stressful events (Deffenbacher et al. 2004), so too can the memory of trained first responders. Indeed, our study primarily explored recognition memory. As shown by Shapiro (2006), recognition tasks may potentially produce greater memory accuracy than free recall or open questions. Therefore, these results may overestimate what officers can remember without cues. Indeed, the free recall narratives showed even fewer details were spontaneously reported by the officers. Similarly, the experience level of our sampled officers may also positively skew memory through, for example, training to attend to certain cues or reducing the amount of anxiety felt during the scenario. The results, therefore, may also overestimate what the "typical" officer (or at least less experienced officers) may be able to remember. It is important to note, therefore, that an expectation of officers to recall specific details correctly may be unrealistic, and leave officers open to suspicion or criticism (as noted above), undermining their legitimacy. These findings will help investigators, the public and media understand that police officers are not immune to the limitations of human memory.

While not a specific hypothesis, the interactions of threat and sensory type were significant. Memory tended to be most accurate for visual threat-relevant details. This may indicate the effect of negative valence on attention (van Steenbergen et al. 2011), with officers more likely to look towards the threat and look longer at threat-related stimuli. While reminiscent of eyewitness "weapon focus", for police officers, this could indicate a more conscious process reflective of their training (to identify and respond to threats). Interestingly, the superiority of recognition of threat over non-threat details was less apparent for spatial/temporal items and actually the reverse for auditory items. Indeed, auditory threat-relevant details were poorly remembered. These largely comprised estimates of the numbers of shots fired by various actors and so echo the findings of Alpert (1987). Although it could be argued that recalling one's own shots includes a behavioural element (trigger pull) as well as auditory, Klinger and Brunson (2009) found that officers who had been involved in a shooting reported auditory exclusion/distortion regarding their own shots, as well as those of others. Our findings show that errors in memory are not

universal across types of detail, and so inaccuracy of detail of one element of an event does not necessarily imply inaccuracy of other, or different forms of, details. In other words, failure of an officer to accurately recall a specific detail should not necessarily undermine their whole account. The findings show the importance of including the sensory type of details in the experimental design when exploring differences in memory for threat and non-threat details. Failing to control for (or counterbalance) the sensory type across conditions could artificially inflate, or mask, differences.

While this study adds to our understanding of officer memory, actual recall after a critical incident is likely to depend on the form and method of questioning, not just the timing. Our findings require further testing exploring police interview practices, including the effects of a "walkthrough" and use of evolving video technology, to better inform policy. Officers in our sample, as noted earlier, also tended to be male, older and more experienced than the typical general duties (patrol) officer. While random assignment to the experimental groups prevented this from affecting the between-group differences (with no significant demographic differences between the groups), replication on a larger and more varied sample would shed light on the extent to which these findings generalise across samples that differ by factors such as age, rank, gender, etc. There is also difficulty in teasing out memory from attention and our study was limited to exploring only the former. Officer-involved shootings can unfold quickly and likely involve multiple stimuli in the environment. It is unlikely officers attend to and process all pieces of available information during this time. When a detail cannot be recalled, or recognised, it is difficult to know whether the point of failure is retrieval or an earlier one of attention or encoding; there may be no optimal interviewing time to retrieve information not attended to and encoded in the first place. It is important to have realistic expectations about what officers likely attend to during these events and, therefore, what information may be retrievable, even under optimal conditions.

In conclusion, despite the limitations of our methodology, our study presents empirical evidence relevant to the important police policy question of when to interview officers after involvement in a critical incident, such as a shooting. Our findings do not show that delaying an interview after involvement in a shooting has any benefits to officers' memory. As Hope (2016) notes, other (non-police) witnesses and victims are typically not afforded a rest period before they are interviewed, since the eyewitness memory literature has shown delay increases the likelihood of forgetting and memory contamination. Notwithstanding obvious injury or trauma, our findings suggest that interviews of police officers should follow the same protocol as for non-police. Delaying an officer interview is likely to worsen memory, particularly for non-threat details of the event. Further, being questioned immediately also helped those officers remember more details when asked again later.

However, it is important to note that it was very common for the officers to misremember details, and also change their responses to questions, showing memory can be unreliable. When officers are involved in critical incidents, they are often expected to recall details of the event, explain their actions and justify their decisions. Where officers are unable to recall details, or their accounts change over time, this can be problematic for an investigation but also potentially undermine the perceived legitimacy of both the officer and the investigation process. Understanding the limitations of officers' memory is, therefore, beneficial not only for investigators, but also the media, public and the courts. Police officers are often held to higher standards than the

general public, but our findings show that, regardless of their specialist training, police officers are still vulnerable to the fallibilities of human memory.

## Appendix: Significant interactions extraneous to hypotheses

While not a hypothesis, the interaction between threat and sensory type was significant for the condition 1 (immediate) and condition 2 (delayed) data, $F(2, 478.34) = 41.60$, $p < 0.001$ (see Fig. 3). Pairwise comparisons showed that participants recognised significantly more visual details that were threat related than non-threat related (mean difference = 28.26, SE = 3.73, $t(1, 478.34) = 7.58$, $p < 0.001$, 95% CI 20.94, 35.59]). Participants also recognised significantly more threat-related spatial/temporal details than non-threat related (mean difference = 8.27, SE = 3.73, $t(1, 478.34) = 2.22$, $p < 0.001$, 95% CI [0.95, 15.60]), although this difference was not as large as for visual details. Conversely, for auditory details, participants recognised significantly fewer threat-related than non-threat-related details (mean difference = − 19.62, SE = 3.73, $t(1, 478.34) = − 5.26$, $p < 0.001$, 95% CI [− 26.95, − 12.29]). The three-way interaction between condition, threat and sensory type was not significant, $F(2, 478.34) = 0.82$, $p = 0.440$.
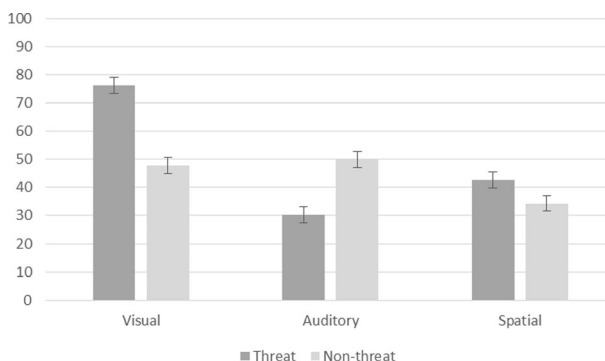
Fig. 3 Mean percentage of correct answers and standard errors, for the two-way interaction of visual, auditory and spatial/temporal details by threat/non-threat (for condition 1 and condition 2)

There was also a significant interaction between threat and sensory type for the condition 1 (immediate) and condition 3 (repeated) data ($F(2, 500.44) = 69.64$, $p < 0.001$) (see Fig. 4). Pairwise comparisons revealed that, when questioned about visual details, participants recognised more threat- than non-threat-related details (mean difference = 24.56, SE = 3.28, $t(1, 500.44) = 7.49$, $p < 0.001$, 95% CI [18.12, 30.99]). However, the opposite was true for auditory details, with less threat than non-threat details recognised correctly (mean

difference = − 29.56, SE = 3.28, $t(1, 500.44) = −9.01$, $p < 0.001$, 95% CI [−36.01, − 23.14]), and there was no significant difference between recognition of threat- and non-threat-related spatial/temporal information (mean difference = 4.03, SE = 3.28, $t(1, 500.44) = 1.23$, $p = 0.22$, 95% CI [−2.41, 10.46).
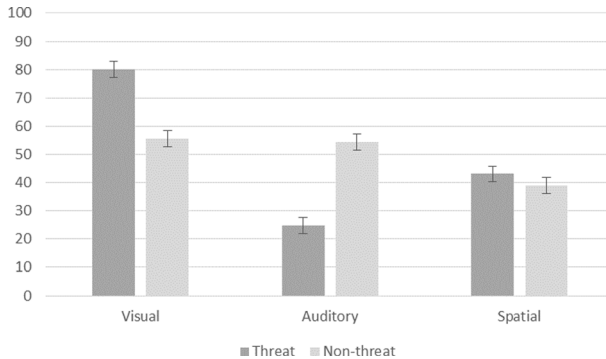
Fig. 4 Mean percentage of correct answers and standard errors, for the two-way interaction of visual, auditory and spatial/temporal details by threat for group 1 (conditions 1 and 3)

The interaction of threat and sensory type was also significant for the condition 2 (delayed) and condition 3 (repeated) data, $(F(2, 470.31) = 43.02, p < 0.001)$ (see Fig. 5). Pairwise comparisons showed that, for visual details, participants recalled significantly more threat-related than non-threat-related items (mean difference = 29.63, SE = 3.74, $t(1, 470.31) = 7.92$, $p < 0.001$, 95% CI [22.28, 36.98]). Participants also recalled significantly more threat-related spatial/temporal details than non-threat related (mean difference = 7.95, SE = 3.74, $t(1, 470.31) = 2.13$, $p = 0.034$, 95% CI [0.60, 15.30]), although this difference was not as large as for visual details. The opposite pattern was evident for auditory details, where participants recalled significantly fewer threat-related than non-threat-related details (mean difference = − 19.30, SE = 3.74, $t(1, 470.31) = −5.16$, $p < 0.001$, 95% CI − 26.65, − 11.96]).
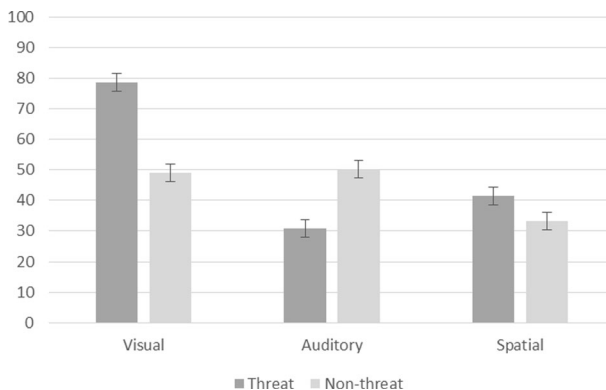
Fig. 5 Mean percentage of correct answers and standard errors, for the two-way interaction of visual, auditory and spatial/temporal details by threat/non-threat (for condition 2 and condition 3)

References

Alferes, V. R., & Kenny, D. A. (2009). SPSS programs for the measurement of nonindependence in standard dyadic designs. *Behavior Research Methods, 41*, 47–54. https://doi.org/10.3758/BRM.41.1.47.

Alpert, G. (1987). Review of deadly force training and policies of the Dallas Police Department: a report to the city of Dallas, Texas. Retrieved 22nd June 2016 from https://www.scribd.com/document/246609261/Alpert-Report-on-Dallas-Police-Deadly-Force

Alpert, G. P., Rivera, J., & Lott, L. (2012). Working toward the truth in officer-involved shootings. *FBI Law Enforcement Bulletin, 81*(5), 1–7.

Armstrong, J., Clare, J., & Plecas, D. (2014). Monitoring the impact of scenario-based use-of-force simulations on police heart rate: Evaluating the royal Canadian mounted police skills refresher program. *Western Criminology Review, 15*(1), 51–59.

Beehr, T., Ivanitskaya, L., Glaser, K., Erofeev, D., & Canali, K. (2004). Working in a violent environment: the accuracy of police officers' reports about shooting incidents. *Journal of Occupational and Organizational Psychology, 77*, 217–235.

Brandt, J., & Benedict, R. H. B. (2001). *Hopkins verbal learning test—revised*. Lutz, FL: Psychological Assessment Resources.

Burke, A., Heuer, F., & Reisberg, D. (1992). Remembering emotional events. *Memory & Cognition, 20*, 277–290.

Christianson, S. Å., Loftus, E., Hoffman, H., & Loftus, G. R. (1990). Eye fixations and accuracy in detail memory of emotional versus neutral events. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 17*, 693–701.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings*. Chicago: Rand McNally.

Deffenbacher, K. A. (1994). Effects of arousal on everyday memory. *Human Performance, 7*, 141–161.

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior, 28*(6), 687–706.

Force Science News #254. n.d. *Force Science Institute details reasons for delaying interviews with OIS survivors*. An interview with William Lewinski. Retrieved from http://www.forcescience.org/fsnews/254.html

Geiselman, R. (2010). Rest and eyewitness memory recall. *American Journal of Forensic Psychology, 28*(2), 1–5.

Genzel, L., Spoormaker, V. I., Konrad, B. N., & Dresler, M. (2015). The role of rapid eye movement sleep for amygdala-related memory processing. *Neurobiology of Learning and Memory, 122*, 110–121. https://doi.org/10.1016/j.nlm.2015.01.008.

Grady, R. H., Butler, B. J., & Loftus, E. F. (2016). What should happen after an officer-involved shooting? Memory concerns in reporting procedures. *Journal of Applied Research in Memory and Cognition, 5*, 246–251. https://doi.org/10.1016/j.jarmac.2016.07.002.

Hartman, M. O., Neill, D., O'Neill, J., & Lewinski, W. (2017). Law enforcement memory of stressful events: recall accuracy as a function of detail type. *Law Enforcement Executive Forum, 17*(3), 22–42.

Hope, L. (2016). Evaluating the effects of stress and fatigue on police officer response and recall: a challenge for research, training, practice and policy. *Journal of Applied Research in Memory and Cognition, 5*, 239–245.

Hope, L., Blocksidge, D., Gabbert, F., Sauer, J. D., Lewinski, W., Mirashi, A., & Atuk, E. (2016). Memory and the operational witness: police officer recall of firearms encounters as a function of active response role. *Law and Human Behavior, 40*, 23–35.

Klinger, D. (2007). *Into the kill zone: a cop's eye view of deadly force*. Hoboken: Josey-Bass.

Klinger, D., & Brunson, R. (2009). Police officers' perceptual distortions during lethal force situations: informing the reasonableness standard. *Criminology and Public Policy, 8*, 117–140.

Lewinski, W., Dysterhefta, J., Priemb, M., & Pettitt, R. (2016). Police officers' actual vs. recalled path of travel in response to a threatening traffic stop scenario. *Police Practice and Research: An International Journal, 17*, 51–67.

Migueles, M., & Garcia-Bajos, E. (1999). Recall, recognition, and confidence patterns in eyewitness testimony. *Applied Cognitive Psychology, 13*(3), 257–268.

Murayama, K., Miyatsu, T., Buchli, D., & Storm, B. C. (2014). Forgetting as a consequence of retrieval: a meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin, 140*, 1383–1409.

Payne, J. D., Schacter, D. L., Propper, R. E., Huang, L. W., Wamsley, E. J., Tucker, M. A., Walker, M. P., & Stickgold, R. (2009). The role of sleep in false memory formation. *Neurobiology of Learning and Memory, 92*, 327–334.

Powell, M. B., & Thomson, D. M. (1997). The effect of an intervening interview on children's ability to remember one occurrence of a repeated event. *Legal and Criminological Psychology, 2*(2), 247–262.

Powell, M., Garry, M., & Brewer, N. (2009). Eyewitness testimony. In I. Freckleton & H. Selby (Eds.), *Expert evidence*. North Ryde: Law Book Co..

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

Shapiro, L. R. (2006). The effects of question type and eyewitness temperament on accuracy and quantity of recall for a simulated misdemeanor crime. *Emporia State Research Studies, 43*(1), 1–7.

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: evolving generalization through selective processing. *Nature Neuroscience, 16*(2), 139–145. https://doi.org/10.1038/nn.3303.

Tombaugh, T. N. (2004). Trail making test a and b: normative data stratified by age and education. *Archives of Clinical Neuropsychology, 19*, 203–214.

Van Steenbergen, H., Band, G. P. H., & Hommel, B. (2011). Threat but not arousal narrows attention: evidence from pupil dilation and saccade control. *Frontiers in Psychology, 2*, 281. https://doi.org/10.3389/fpsyg.2011.00281.

Weisburd, D., & Britt, C. (2007). *Statistics in criminal justice* (3rd ed.). New York: Springer.

Weschler, D. (1945). A standardized memory scale for clinical use. *The Journal of Psychology, 19*, 87–95.

White, M. D., Ready, J. T., Kane, R. J., Yamashiro, C. T., Goldsworthy, S., & Bonds McClain, D. (2015). Examining cognitive functioning following TASER exposure: a randomized controlled trial. *Applied Cognitive Psychology, 29*, 600–607.

Wolchover, D., Heaton-Armstrong, A., Hope, L., & Gabbert, F. (2014). Notebook challenge: no conferring—part III. *Criminal Law and Justice Weekly, 178*, 266–268.

Louise E. Porter is a Senior Lecturer in the School of Criminology and Criminal Justice at Griffith University, Australia, and member of the Griffith Criminology Institute. She has a PhD and Masters in Investigative Psychology from the University of Liverpool, UK. Her research on policing focuses on police misconduct and use of force, as well as police oversight and integrity management. At Griffith, she teaches in the areas of professional ethics and corruption prevention and the intersection between psychology and the justice system.

Justin Ready joined the School of Criminology and Criminal Justice at Griffith University as a Senior Lecturer in 2016. He received his PhD at Rutgers University, where he served as Project Director in the Centre for Crime Prevention Studies. He later worked as a Senior Research Associate at the Police Foundation in Washington, DC. He teaches courses in research methods, theory and statistics. His research focuses on crime hot spots and the impact of new technology on police practices. He is the 2013 recipient of the Young Experimental Scholar Award and a 2016 recipient of the Award for the Outstanding Experimental Field Trial. His publications have appeared in peer-reviewed journals such as *Criminology*, *Justice Quarterly* and the *Journal of Experimental Criminology*.

Geoffrey P. Alpert is a Professor of Criminology and Criminal Justice at the University of South Carolina and has an appointment at Griffith University. He has taught at the FBI National Academy, the Federal Law Enforcement Training Center and The Senior Management Institute for Police. He is currently a Federal Monitor for the New Orleans Police Department and on the compliance team for the Portland, Oregon Police Bureau. He has testified to Congress, several state legislatures and the President's Task Force on 21st Century Policing and is a member of the International Association of Chiefs of Police Research Advisory Committee and the PERF Research Advisory Board. For the past 30 years, his research interests have focused on police use of force, emergency driving and the linkages between researchers and practitioners.