Cognitive and metacognitive determinants of eyewitness memory accuracy over time

Anat Adi Shapira and Ainat Pansky

Department of Psychology and the Institute of Information Processing and Decision Making

University of Haifa

Please address correspondence to:

Ainat Pansky

Department of Psychology

University of Haifa

199 Aba Khoushy Ave.

Haifa 3498838

Israel

e-mail: pansky@research.haifa.ac.il

TEL: +972-4-8249437

FAX: +972-4-8249431

Running head: Eyewitness memory accuracy over time

Compliance with ethical standards:

Funding: This research was supported by The Israel Science Foundation (Grant 819/13) awarded to Ainat Pansky.

Conflict of interest: The authors declare that they have no conflict of interest.

Abstract

In the present study, we investigated the accuracy of eyewitness accounts over time from a metacognitive perspective, in which post-retrieval monitoring and control processes play a crucial role in mediating between memory retrieval and ultimate memory performance. In two experiments, participants viewed a narrated slide show depicting ordinary daily events and were questioned about fine-grained event details, either immediately or after a delay (of either 24 or 48 hours). High motivation for accurate responding was induced via monetary incentives (Exp. 1) or instructions (Exp. 2). Using Koriat and Goldsmith's (1996) *Quantity-Accuracy Profile* methodology, we were able to isolate the cognitive and metacognitive components posited to underly free-report memory accuracy, and to examine them over time. Our results showed that, even under conditions of free-report and high motivation for accurate responding, the accuracy of memory reports declined substantially over time, largely due to reduced monitoring effectiveness (i.e., monitoring resolution) rather than to changes in control policy (i.e., report criterion). As predicted, the decline over time in confidence was more pronounced for true than for false memories, such that the metacognitive ability to differentiate between correct and incorrect answers deteriorated with time. This poorer monitoring resolution resulted in an increased proportion of errors among the volunteered responses, and consequently, in lower free-report accuracy. Our results shed light on the manner in which memory accuracy over time is affected by changes in the effectiveness of the metacognitive processes that operate during memory reporting.

Cognitive and metacognitive determinants of eyewitness memory accuracy over time

Consider a situation in which two individuals are interrogated by the police after witnessing a crime. They are asked the same series of questions about fine-grained critical event details, but one eyewitness is interviewed immediately after the event whereas the other is interviewed 48 hours later. Both eyewitnesses are requested to be accurate on every single question and to refrain from providing answers they are unsure about. Which of the two witness accounts would you put more faith in and why?

Over the past decades, there has been a growing awareness of the substantial fallibility of memory and its potential consequences in everyday life in general, and in the eyewitness context in particular (for reviews, see Koriat, Goldsmith, & Pansky, 2000; Loftus, 2003). One of the reasons why eyewitness accounts might be incomplete and/or unreliable is that they are often collected a considerable time after the initial incident (e.g., Hope, Gabbert, & Fisher, 2011). Limited resources often restrict opportunities to interview witnesses for several days or even weeks after the incident, particularly if they are not directly implicated. Delaying retrieval from memory has been systematically found to decrease the amount of information that can be recollected (e.g., Penrod, Loftus, & Winkler, 1982; Tuckey & Brewer, 2003). Over time, access to detailed information is selectively impaired, as fine-grained, verbatim-level information decays more rapidly than coarse-grained, gist-level information (e.g., Goldsmith, Koriat, & Pansky, 2005; Koriat, Levy-Sadot, Edry, & de Marcas, 2003). Obviously, eliciting complete, accurate, and detailed reports from eyewitnesses is critical in many different contexts, including the investigation of a crime, occupational accidents, or security incidents. In the present study, we investigated the reliability (i.e., accuracy) and completeness (i.e., quantity) of eyewitness accounts over time from a metacognitive perspective, in which monitoring and control processes play a crucial role in mediating

between memory retrieval and ultimate memory performance (see Goldsmith & Koriat, 2008).

Memory quantity vs. memory accuracy over time

As suggested by Koriat and Goldsmith (1994, 1996), in assessing free-report performance, one can distinguish between two properties of memory. I*nput-bound quantity* reflects the likelihood that each event detail (i.e., input item) is correctly remembered, whereas *output-bound accuracy* reflects the conditional probability that each reported event detail is correct. Clearly, obtaining as much as possible event information is critical for solving crimes, but it is at least equally important that this information can be relied upon to be faithful to the event in order to prevent miscarriages of justice. The courtroom oath "to tell the whole truth and nothing but the truth" demonstrates the requirement that eyewitnesses simultaneously uphold these two memory goals: to provide as much veridical information as possible and to avoid the reporting of erroneous information. If memory monitoring were perfect, an eyewitness should be able to distinguish between correct and incorrect information that comes to mind and achieve optimal quantity and accuracy. However, numerous studies have shown that memory monitoring is imperfect (e.g., Benjamin & Bjork, 1996; Johnson, Hashtroudi, & Lindsay, 1993; for a review, see Pansky, Koriat, & Goldsmith, 2005), such that improving accuracy typically comes at the expense of quantity, leading to a typical reduction in the amount of correct reported information, in what is known as the quantity-accuracy tradeoff (e.g., Koriat & Goldsmith, 1994, 1996).

Many studies have shown that the amount of event details that can be recollected declines with time (e.g., Pansky, 2012; Pansky, Tenenboim, & Bar, 2011). Following the pioneering work of Ebbinghaus (1895/1964), subsequent research has confirmed that the course of forgetting is a curvilinear function of retention interval, with a relatively large initial decline in memory quantity and decreasing additional declines thereafter (for a review,

see Rubin & Wenzel, 1996). In contrast to this well-documented decrease in memory quantity over time, the findings regarding memory accuracy are mixed. Whereas some studies have demonstrated a decrease in output-bound accuracy (e.g., Bahrick, Hall, & Dunlosky, 1993; Bergman & Roediger, 1999; Koriat, Goldsmith, Schneider, & Nakash-Dura, 2001; Larsson, Granhag, & Spjut, 2003), other studies have found unexpectedly stable accuracy rates for memories of events over time (e.g., Ebbesen & Rienick, 1998; Evans & Fisher, 2011). In the present study, we examined memory accuracy over time within a conceptual framework specifying the mediating role of cognitive and metacognitive processes.

Metacognitive monitoring and control processes under free-report conditions

Koriat and Goldsmith (1996) have put forward a model of the strategic regulation of memory reporting under free-report conditions, in which post-retrieval metacognitive monitoring and control processes play a crucial role. According to the model, rememberers use a monitoring mechanism to subjectively assess the correctness of potential memory responses, and a control mechanism then determines whether or not to volunteer the best accessible candidate answer. The control mechanism operates by setting a report criterion on the monitoring output: The answer is volunteered if its assessed probability of being correct passes the criterion, but is withheld otherwise. The report criterion is set on the basis of implicit or explicit payoffs: the perceived gain for providing correct information relative to the cost of providing wrong information. In empirical experiments accuracy motivation is usually manipulated by using a payoff matrix which offers a monetary bonus for each correct volunteered answer and a penalty for each incorrect volunteered answer (e.g., Kelley & Sahakyan, 2003; Koriat & Goldsmith, 1994, 1996; Pansky & Goldsmith, 2014; Pansky, Goldsmith, Koriat, & Pearlman-Avnion, 2009; see also McCallum, Brewer, & Weber, 2016).

In order to isolate and measure the cognitive and metacognitive components posited to underly free-report quantity and accuracy performance, Koriat and Goldsmith (1996) have developed an experimental paradigm and assessment methodology, *Quantity-Accuracy Profile*, that includes both free and forced reporting along with the elicitation of confidence judgments. Results from several studies (both simulation analyses and empirical results) have provided strong support for the model, revealing the manner in which post-retrieval monitoring and control processes mediate between memory retrieval and ultimate memory performance (e.g., Goldsmith et al., 2005; Higham, 2002, 2007; Higham & Tam, 2005; Koriat & Goldsmith, 1994, 1996; Kelley & Sahakyan, 2003; Pansky et al., 2009; Rhodes & Kelley, 2005; for a review see Goldsmith & Koriat, 2008). These studies have shown that when no deliberate attempt to impair memory monitoring is made, there is an option to decide which items to report, and accuracy motivation is high, rememberers are able to enhance their memory accuracy substantially by screening out answers that they consider likely to be wrong (see also Koriat et al., 2001; Roebers, Moga, & Schneider, 2001; Roebers & Schneider, 2005). Importantly, the ability to improve free-report accuracy at a relatively small cost in terms of the quantity of correct reported information depends upon the following: (a) reasonable success in monitoring the correctness of candidate answers that come to mind (i.e., effective monitoring), (b) high *control sensitivity* (i.e., heavily relying on subjective confidence in deciding whether to volunteer or withhold an answer), and (c) setting an appropriate *control policy* (i.e., report criterion) for the specific reporting context. Control sensitivity has been found to be at ceiling level among young healthy adults (e.g., Koriat & Goldsmith, 1996; Kelley & Sahakyan, 2003; Pansky, et al., 2009; Rhodes & Kelley, 2005), and the report criterion tends to be relatively high under conditions of high-accuracy motivation. Therefore, the effectiveness of the monitoring process plays an important role in affecting memory

accuracy performance. When the ability to distinguish between correct and incorrect candidate answers improves, greater increases in accuracy can be achieved at lower costs in quantity. In contrast, when monitoring effectiveness is poor, the exercise of report option may yield little or no benefit in accuracy, and might merely reduce the quantity of the reported information. For typical (moderate to high) levels of monitoring effectiveness, enhancing accuracy becomes relatively costly in terms of quantity performance as the criterion level is raised (e.g., Goldsmith, 2016; Koriat & Goldsmith, 1994, 1996; Koriat et al., 2001).

Given the documented ability of rememberers to strategically regulate memory accuracy, how can one account for findings of declining accuracy over time, even under conditions of free-report, normal-monitoring, and high accuracy motivation (e.g., Bahrick et al., 1993; Bergman & Roediger, 1999; Koriat et al., 2001; Larsson et al., 2003)? Despite the typical decline in memory quantity over time, free-report accuracy would be expected to remain stable if no changes occurred over time in the metacognitive components. That is, if the monitoring and control processes continued to operate in the same manner, accuracy-motivated rememberers would be expected to freely report less information over time, but to maintain the accuracy of the reported information.

Various observations in the literature raise two tentative hypotheses as to the metacognitive changes that might occur over time and could account for declining accuracy. First, even soon after exposure to an event, rememberers are likely to recollect some wrong event details: Prior knowledge, expectations, and schemas (e.g., Bartlett, 1932; Neisser, 1996) guide reproductive, (re-)constructive and deductive processes that tend to bring to mind prototypical, gist-consistent, or schema-consistent information that is sometimes incorrect (e.g., Brewer & Treyens, 1981; Pansky & Tenenboim, 2011; Schacter, Guerin, & St. Jacques, 2011; Tuckey & Brewer, 2003). Despite their inaccuracy, such false memories

are often held with relatively high confidence (e.g., Garcia-Bajos & Migueles, 2003; Pansky & Koriat, 2004). A well-established finding termed *false memory persistence* is that false memories exhibit a milder rate of decline over time than true memories, under both free-report and forced-report conditions (e.g., Brainerd , Reyna, & Brandse, 1995; McDermott, 1996; Payne, Elie, Blackwell, & Neuschatz, 1996; Seamon et al., 2002; Thapar & McDermott, 2001). A prominent account of false memory persistence was suggested in the context of fuzzy-trace theory (e.g., Brainerd & Reyna, 1993), according to which, each studied item is encoded at various levels of precision, from verbatim traces representing detailed episodic information to gist traces capturing its meaning. Over time, verbatim traces become inaccessible more rapidly than gist traces (e.g., Brainerd & Reyna, 1998). Assuming that memory for the studied information is supported mainly by verbatim traces and false memory is supported mainly by gist traces, true memories are less resistant to forgetting than false memories, resulting in false memory persistence (Brainerd & Reyna, 2002). Confidence in such false memories has also been shown to be more stable over time than confidence in true memories (e.g., Garcia-Bajos & Migueles, 2003; Toglia, Neuschatz, & Goodwin, 1999; Weinstein, McDermott, & Chan, 2010). Thus, over time, the confidence judgments might become less effective in discriminating between correct and incorrect information that comes to mind (i.e., lower monitoring effectiveness). Consequently, less high-confidence correct responses might persist over time than high-confidence errors, resulting in an increasing proportion of errors among the volunteered responses and lower output-bound accuracy over time. Second, as suggested by Ackerman and Goldsmith (2008), respondents are expected to provide substantive answers to at least some of the questions that they are asked. In situations in which relatively little is remembered, such as after a delay, respondents may be reluctant to say “I don’t know” (i.e., to withhold their responses) too often, due to personal-social expectations for informativeness and cooperation, and might adopt a more liberal report

criterion (see also Kelley & Sahakyan, 2003; Pansky et al., 2009). Such lowering of the report criterion, at delayed testing, is likely to yield a decrease in accuracy (e.g., Koriat & Goldsmith, 1996).

To test these hypotheses, we examined whether the anticipated inferior quality of eyewitness reports at delayed testing would result from poorer monitoring effectiveness, changes in control policy (i.e., an adoption of an overly liberal report criterion), or both. As we were primarily interested in free-report memory performance and aimed to examine it as uncontaminated as possible, we used a version of Koriat and Goldsmith's (1996) methodology by which the free-report test was conducted before the forced-report test. This order of the test formats, compared to the reverse order (forced-report before free-report) has been recently shown to yield more accurate free-report performance with no cost to memory quantity (Hollins & Weber, 2017). In the free-report phase, for each cued-recall question, the participants were allowed to either report a fine-grained event detail or refrain from reporting it, under high accuracy motivation conditions. Each answer was followed by a confidence judgment estimating the subjective likelihood of its correctness. In the forced-report phase, the participants were required to provide their best guess answer to each question they had chosen to leave unanswered in the previous phase, and to provide a confidence judgment.

This procedure allowed the derivation of a rich profile of measures:

1. Forced-report quantity–the proportion of correct answers provided in both phases out of the total number of test questions, as an estimate of the amount of information that is accessible in memory.
2. Confidence–the mean confidence assigned to the responses. Confidence was converted to an assessed probability of correctness ranging between 0 and 1 by dividing each confidence judgment by 100, in order to allow its comparison to actual proportion correct.

3. Volunteering rate–the proportion of responses (whether correct or incorrect) that were freely-reported out of the total number of test questions.

4. Monitoring effectiveness–the correspondence between the correctness of the answers and the confidence associated with them (see Lichtenstein, Fischhoff, & Phillips, 1982; Schraw, 2009), was indexed in terms of: (a) absolute correspondence, assessed by calibration bias scores—the extent to which mean confidence was higher (over-confidence) or lower (under-confidence) than mean accuracy, and (b) relative correpondence—the extent to which the subjective confidence judgments successfully distinguished between correct and incorrect answers, assessed by two measures of monitoring resolution: (1) The adjusted normalized discrimination index (ANDI, see Yaniv, Yates, & Smith, 1991, for a detailed description of how ANDI is calculated), and (2) a simple discrimination index (see Schraw, 2009). ANDI ranges from 0 (no discrimination) to 1 (perfect discrimination), and reflects the amount of variance in accuracy accounted for by participants' confidence ratings. Thus, for example, an ANDI score of .40 indicates that the confidence judgments can explain 40% of the variability in accuracy. We chose ANDI as a measure of monitoring resolution due to its two advantages over other measures: (1) ANDI is normalized in terms of variance in the outcome (i.e., accuracy), such that its interpretation is not conditional on the objective uncertainty of the predicted outcome, and (2) ANDI is unaffected by the number of judgments in different confidence categories (Yaniv et al., 1991). Another reason for choosing ANDI is that problems were recently identified with regard to the most common measure of resolution used in the metacognitive literature, the Goodman-Kruskal Gamma coefficient (Goodman & Kruskal, 1954). For example, Gamma was criticised for being affected by response bias (independent of discrimination skill; see, e.g., Benjamin & Diaz, 2008; Masson & Rotello, 2009). In addition to ANDI, we also calculated a simple discrimination index—the difference between mean confidence for correct responses and mean confidence for incorrect responses

(termed ‘the slope’ in Ronis & Yates, 1987; see also Allwood, Innes-Ker, Homgren, & Fredin, 2008; Stankov & Crawford, 1996). The simple discrimination index was chosen due to its sensitivity to the relative decline over time in confidence for correct versus incorrect answers, as well as its simple, straightforward interpretation.

5. Free-report accuracy—the proportion of correct answers out of the total number of answers volunteered in the free-report phase.

6. Free-report quantity—the proportion of correct answers volunteered in the free-report phase out of the total number of test questions.

7. Report criterion—the minimal level of confidence that was required by each participant in order to volunteer an answer, estimated as the cut-off point in confidence that best separated between the items that were volunteered and those that were withheld in phase I.

In addition to replicating the well-documented decline over time in forced- and free-report memory quantity (e.g., Rubin & Wenzel, 1996), our novel predictions were that:

1. This decline would be accompanied by declines in: (a) mean confidence, (b) volunteering rate, (c) monitoring effectiveness, (d) free-report accuracy, and (e) report criterion.

2. The decline over time in confidence and volunteering rate would be more pronounced for true memories than for false memories.

As mentioned above, according to Koriat and Goldsmith (1996) model, free-report memory performance is influenced by one's report criterion, which is set according to the perceived gain for providing correct information relative to the cost of providing wrong information. In Experiment 1, accuracy motivation was induced by using explicit monetary incentives, as typically done in experimental contexts (e.g., Higham, 2002, 2007; Higham & Tam, 2005; Kelley & Sahakyan, 2003; Koriat & Goldsmith, 1996; Koriat et al., 2001; Pansky et al., 2009; Pansky & Goldsmith, 2014; Pansky & Nemets, 2012; Portnoy & Pansky, 2016; Rhodes & Kelly, 2005). Keeping in mind that in forensic interrogations, eyewitnesses are not

given monetary or point incentives for accurate responding, one of the goals of the present study was to examine the effectiveness of a more ecologically-valid manipulation. Toward this end, accurate responding in the free-report phase was induced by using accuracy-motivating instructions in Experiment 2.

Experiment 1

In Experiment 1, the participants initially viewed a narrated slide show containing ten concrete target items. Because our primary interest was to examine the contribution of both retrieval per se and monitoring and control processes to free-report memory performance over time, we chose event details that were likely to fade away with time. Previous research has shown that details that were peripheral to the meaning of the event were more vulnerable to the effects of time than more central details (e.g., Christianson & Loftus, 1987; Flowe, Takarangi, Humphries, & Wright, 2016), although both types of details can be of critical importance with regard to the guilt of innocence of a suspect (see Read & Connolly, 2007). Therefore, the target items we chose were not thematically central to the narrative of the slide show. However, all the target items were clearly visible, and each was presented on a separate slide. Preliminary perceptual testing showed that participants (n = 20) who were questioned about each target item while viewing the relevant slide correctly answered 99.5% of the questions, and all of them stated that they had no difficulty answering the target questions based on what they observed in the slides. Preliminary memory testing (n = 20) confirmed that our target items were of varied difficulty (with forced-report proportion correct ranging between .30 and 1.00 at immediate testing, M = .54, SD = .50), and that all of the items had the potential to undergo forgetting over time.

Previous research on the effects of retention interval on memory quantity have shown that there is a gradual decline in memory performance over time and that this decline is curvilinear rather than linear (for a review, see Wixted, 2004). To see whether we would

replicate this pattern for free-report accuracy as well, we linearly increased the retention interval by 24 hours in two steps, conducting memory testing either immediately, after a 24-hour delay, or after a 48-hour delay. Earlier research with similar materials (Pansky & Nemets, 2012) has shown significant decline in forced-report quantity correct over the entire 48-hour interval. However, how gradual a decline would be found across this retention interval with regard to both cognitive and metacognitive measures remained to be examined.

Method

Participants and Design. Sixty native Hebrew-speaking students (27% males and 73% females, mean age= 24.33) from the University of Haifa took part in the experiment. They were randomly and equally assigned to the immediate, 24-hour, or 48-hour retention interval groups, with 20 participants in each group. Sample size was determined based on power calculations performed with G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) conducted on preliminary data we collected. These calculations indicated that the sample size required for detecting differences with 80% power at $\alpha = .05$ should exceed 19 participants in each group. The participants were paid 30 NIS (approximately 8 US$) for their participation in each experimental session (one session in the case of the immediate-testing group, and two sessions in the case of the delayed-testing groups).

Materials. A 6.5-min computerized slide show was used as the target event. The slide show, developed by Pansky et al. (2011), consisted of 33 still color pictures accompanied by a corresponding narration, telling a story about a day in a female student's life. Each slide was presented for 10.1 s. Ten concrete items (e.g., Carlsberg beer), each presented on a separate slide and clearly visible, constituted the target items (see Appendix A, column 2, in the supplementary material). The rest of the experiment was run using a computer program developed with E-Prime experiment-generating software. The memory test consisted of 10 cued-recall questions, each referring to one of the target items (e.g.,

"Which beer did Itai order for himself, when he was chatting with Inbal at the pub? (Please name the beer brand)." For the full set of questions, see Appendix A, column 3. The order of the questions on the memory test corresponded to the chronological order in which the target items were presented in the slide show.

Procedure. The experiment was administered in small groups of one to six participants. In the first stage of the experiment, the participants were informed that they would view a slide show and would then answer questions about its content. After viewing the slide show, the participants performed a non-verbal filler task of solving Raven's Progressive Matrices for approximately ten minutes. One third of the participants proceeded immediately to the second stage in which the cued-recall test was administered, whereas the other two groups of participants completed this stage in a separate session, after either 24 or 48 hours. The entire experiment, whether separated or not, lasted about 40 minutes. For each of the ten questions, the participants were required to provide a fine-grained answer using the following two-phase free-forced procedure (e.g., Koriat & Goldsmith, 1996). In the initial free-report phase, in response to each cued-recall question, the participants were allowed to either report the event detail or refrain from reporting it, under high accuracy motivation conditions. If an answer was volunteered, a confidence judgment was also solicited, estimating the likelihood that the answer was correct (on a 0-100% scale). The participants were told that a confidence judgment of 0% indicates a total lack of confidence in the correctness of an answer, whereas a confidence judgment of 100% indicates complete certainty in its correctness. They were asked to use the entire range of the confidence scale, rather than a limited number of values. Accurate responding in the free-report phase was encouraged using an explicit high-incentive payoff schedule: The participants were paid 2 NIS (approximately US$ 0.50) for each volunteered correct answer, and penalized 5 NIS (approximately US$ 1.25) for each volunteered incorrect answer. They were told that they

would neither be penalized nor would receive a bonus for withheld responses and that although they might not break even, they would not have to pay any losses. In the forced-report phase, questions that were not answered in the previous phase were repeated and the participants were required to provide an answer to each, followed by a confidence judgment. No monetary incentive was offered in the forced-report phase. See Appendix B (in the supplementary material) for the instructions that were presented to the participants at the various stages of Experiment 1.

Results and Discussion

Two independent judges determined for each response that was provided whether it was correct or incorrect. Responses were classified as correct only if they completely matched the target items (listed in Appendix A in the supplementary material) at the fine-grained level and constituted valid responses to the precise questions that were asked. The classifications made by these two judges were identical in 99.67% of the cases. A third judge determined the scoring of the controversial .33% of the responses.

Each of these dependent variables was subjected to an ANOVA, with retention interval as the between-subject factor. Follow-up pairwise comparisons (using t-tests) were conducted between immediate and 24-hour testing, and between 24-hour and 48-hour testing.

Forced-report quantity. As shown in Figure 1 (panel A), forced-report proportion correct declined across the three retention intervals, $F(2, 57) = 7.78$, $p = .001$, $\eta^2_p = .22$. Pairwise comparisons revealed that the relatively large initial decline (of .14) over the first 24 hours was significant, $t(38) = 2.94$, $p = .006$, $d = .95$, whereas the smaller decline (of .04) over the next 24 hours was not, $t(38) = .82$, $p = .417$, $d = .27$.

< Insert Figure 1 about here >

Confidence. As also shown in Figure 1 (panel A), the participants' confidence in the correctness of their answers also declined over time, $F(2, 57) = 8.24$, $p = .001$, $\eta^2_p = .22$.

Here too, the initial decline (of .12) over the first 24 hours was significant, $t(38) = 2.65$, $p = .012$, $d = .86$, whereas the decline (of .09) over the next 24 hours was not, $t(38) = 1.47$, $p = .150$, $d = .48$.

Volunteering rate. As additionally shown in Figure 1 (panel A), the proportion of the responses that were volunteered in the free-report phase, out of the ten test questions, averaged .72, .59, and .58 at immediate, 24-hour, and 48-hour testing, respectively. Although the pattern of decline resembled that of the confidence data, the numerical decline in volunteering rate over time did not reach statistical significance, $F(2, 57) = 2.14$, $p = .127$, $\eta^2_p = .07$. To further explore the somewhat different pattern of decline of the volunteering rates compared to the confidence judgments, we conducted an additional ANOVA. In this analysis, the difference between confidence and volunteering rate was treated as a repeated factor, for the assessment of a possible interaction between this factor and retention interval. However, the interaction was not statistically significant, $F(2, 57) = 1.65$, $p = .201$, $\eta^2_p = .06$.

Over the three retention intervals, the confidence judgments were highly diagnostic of the volunteering of the answers, averaging .70 for volunteered answers and .21 for withheld answers, $F(1, 51) = 354.90$, $p < .001$, $\eta^2_p = .87$. The difference between the confidence judgments of the volunteered and withheld responses was substantial and significant for all three retention intervals [.60 at immediate testing, $t(16) = 12.52$, $p < .001$, $d = 4.29$, .47 after 24 hours, $t(18) = 10.23$, $p < .001$, $d = 3.11$, and .41 after 48 hours, $t(17) = 9.76$, $p < .001$, $d = 2.26$]. However, this difference did decrease over time, with a significant interaction between retention interval and volunteering, $F(2, 51) = 4.42$, $p = .017$, $\eta^2_p = .15$. This was due to a significant decline over time (from .81 to .61) in the confidence for volunteered responses, $F(2, 58) = 7.95$, $p = .001$, $\eta^2_p = .221$, without a parallel decline in the confidence for withheld answers, $F(2, 54) = .405$, $p = .669$, $\eta^2_p = .015$.

Was there a significant deterioration in monitoring effectiveness over time, as predicted? Did the absolute and relative correspondence between the correctness of the participants' answers and the confidence associated with them decline or did it remain stable over the 48-hour retention period?

Calibration bias. To assess the degree of over/under-confidence, we conducted a mixed-model ANOVA, in which the discrepancy between predicted performance (i.e., confidence) and actual performance (i.e., forced-report quantity) was treated as a repeated factor. This treatment allowed the assessment of a possible interaction between calibration bias and retention interval. Although the participants were found to be somewhat over-confident in the correctness of their responses, $F(1, 57) = 10.87$, $p = .002$, $\eta^2_p = .16$, the degree of this calibration bias did not change over time, $F(2, 57) = .44$, $p = .644$, $\eta^2_p = .02$.

The adjusted normalized discrimination index (ANDI). Using ANDI as a measure of monitoring resolution, we found a deterioration over time in the ability of the participants to differentiate correct from incorrect information that came to mind, $F(1, 57) = 3.10$, $p = .053$, $\eta^2_p = .16$[1] (see Figure 2, panel A). There was a significant decline (of .20) in ANDI over the first 24 hours, $t(38) = 2.38$, $p = .022$, $d = .77$, but no significant change over the next 24 hours, $t(19) = .59$, $p = .620$, $d = .19$.

< Insert Figure 2 about here >

Discrimination index. As a second measure of monitoring resolution, we compared the mean confidence for correct versus incorrect answers. Over the three retention intervals, the confidence judgments were diagnostic of the correctness of the answers, averaging .67 for correct answers and .40 for incorrect answers (discrimination index = .27), $F(1, 57) = 146.41$,

[1] Note, though, that ANDI was significantly different from zero in each of the three retention intervals: $t(19) = 5.82$, $p < .001$, $d = 2.67$, $t(19) = 3.74$, $p = .001$, $d = 1.72$, and $t(19) = 3.74$, $p = .001$, $d = 1.72$, at immediate, 24-hour, and 48-hour testing respectively, suggesting that the participants were able to discriminate between correct and incorrect information even at delayed testing.

$p < .001$, $\eta^2_p = .73$. However, as with ANDI, the ability to differentiate correct from incorrect answers deteriorated over time, with a significant interaction between correctness and retention interval, $F(1, 57) = 5.50$, $p = .007$, $\eta^2_p = .16$[2]. Examination of two retention intervals each time revealed a significant interaction between correctness and retention interval when comparing immediate to 24-hour testing, $F(1, 38) = 4.83$, $p = .034$, $\eta^2_p = .11$, but no interaction when comparing 24-hour to 48-hour testing, $F(1, 38) = 1.23$, $p = .274$, $\eta^2_p = .03$. Thus, both resolution measures showed that the metacognitive ability to discriminate between correct and incorrect answers that came to mind deteriorated over the initial 24 hours, such that the confidence judgments became less diagnostic of correctness, with no further decline in monitoring resolution thereafter. This was a result of a differential effect of retention interval on confidence in true and false memories, as will be shown next. The correlation between the discrimination index and ANDI was .65 ($p < .001$), a strong correlation one would expect from two measures of monitoring resolution, but not too perfect a correlation to render one of them redundant.

< Insert Figure 3 about here >

As predicted and as can be seen in Figure 3 (panel A), the confidence judgments for correct answers declined significantly over time, averaging .80, .67, and .55 at immediate, 24-hour, and 48-hour testing, respectively, $F(2, 57) = 7.62$, $p = .001$, $\eta^2_p = .21$. Pairwise comparisons revealed that the decline over the first 24 hours was significant, $t(38) = 2.27$, $p = .029$, $d = .74$, but the decline over the next 24 hours was not, $t(38) = 1.60$, $p = .118$, $d = .52$. In contrast, the confidence judgments for incorrect answers did not show a significant decline

[2] Despite this decline in the discrimination index over time, the confidence judgments did discriminate between correct and incorrect responses in each of the three retention intervals separately: at immediate testing (discrimination index = .37), $t(19) = 8.45$, $p < .001$, $d = 3.88$; after 24 hours (discrimination index = .25), $t(19) = 7.40$, $p < .001$, $d = 3.40$, and after 48 hours (discrimination index = .19), $t(19) = 5.06$, $p < .001$, $d = 2.32$.

over time, averaging .43, .42, and .36 at immediate, 24-hour, and 48-hour testing, respectively, $F(2,57) = .90$, $p = .411$, $\eta^2_p = .03$.

Number of volunteered correct vs. incorrect answers. Did this phenomenon of declining confidence in true memories versus stability in confidence in false memories influence the behavior of the participants in terms of the number of correct versus incorrect answers that they chose to volunteer? Indeed, although more correct than incorrect answers were volunteered overall, $F(1, 56) = 14.65$, $p < .001$, $\eta^2_p = .21$, this pattern changed over time, $F(2, 56) = 11.07$, $p < .001$, $\eta^2_p = .28$. Thus, the discrimination between correct and incorrect answers in terms of the participants' volunteering decisions deteriorated one day after witnessing the target event, $F(1, 38) = 12.59$, $p = .001$, $\eta^2_p = .25$, resulting in no discrimination whatsoever at 24-hour and 48-hour delayed testing, $F(1, 37) = .27$, $p = .604$, $\eta^2_p = .01$, and $F(1, 37) = 1.37$, $p = .240$, $\eta^2_p = .04$, respectively.

< Insert Figure 4 about here >

As shown in Figure 4 (panel A), the number of volunteered correct answers declined significantly over time, $F(2, 56) = 11.85$, $p < .001$, $\eta^2_p = .30$, averaging 5.05, 3.20, and 2.95 at immediate, 24-hour, and 48-hour testing, respectively, and mirroring the decline in the confidence judgments for correct answers. Pairwise comparisons revealed that the decline over the first 24 hours was significant, $t(38) = 4.19$, $p < .001$, $d = 1.36$, with no further decline over the next 24 hours, $t(38) = .53$, $p = .600$, $d = .17$. In contrast, the number of volunteered incorrect answers did not decline across the three retention intervals, actually showing a slight increase over time which was not statistically significant, $F(2, 56) = 2.03$, $p = .141$, $\eta^2_p = .30$, averaging 2.10, 2.65, and 3.16 at immediate, 24-hour, and 48-hour testing, respectively. Thus, false memory persistence was exhibited in the volunteering decisions, with a significant decline over time in the number of volunteered correct answers but no decline in the number of incorrect answers the participants chose to freely report.

To what extent did the decline over time in the volunteering rate of correct answers but not in the volunteering rate of incorrect answers affect free-report memory performance? To answer this question, we examined free-report memory accuracy and quantity performance. These analyses are based on 59 participants, because one participant in the delayed 48-hour group volunteered no answers in the free-report phase.

Free-report accuracy. As shown in Figure 5 (panel A), free-report accuracy declined across the three retention intervals, $F(2, 56) = 8.62$, $p = .001$, $\eta^2_p = .24$, with a non-significant decline (of .12) over the first 24 hours, $t(38) = 1.97$, $p = .056$, $d = .64$, and a significant decline (of .13) over the next 24 hours, $t(37) = 2.11$, $p = .042$, $d = .69$.

< Insert Figure 5 about here >

Free-report quantity. Free-report quantity declined over time as well (see Figure 5, panel A), $F(2, 56) = 11.85$, $p < .001$, $\eta^2_p = .30$, with a significant decline (of .19) over the first 24 hours, $t(38) = 4.19$, $p < .001$, $d = 1.36$, but no further decline over the next 24 hours, $t(37) = .53$, $p = .600$, $d = .17$.

Quantity-accuracy tradeoff. We next examined whether there was a change over time in the extent to which the gain in accuracy achieved through free reporting came at a cost in terms of quantity (i.e., the quantity-accuracy tradeoff). Based only on the decline in monitoring resolution that we found over time, one might expect to find a larger quantity-accuracy tradeoff at delayed testing. However, this tradeoff is actually a product of a complex interplay between the overall level of retention, monitoring effectiveness, and the report criterion (see Koriat & Goldsmith, 1996), and is therefore very difficult to predict. Comparing free-report to forced-report performance, the option of free-report allowed the participants to achieve a gain (of .14) in accuracy, $F(1, 56) = 59.29$, $p < .001$, $\eta^2_p = .51$, which did not significantly change over time, $F(2, 56) = 2.85$, $p = .066$, $\eta^2_p = .09$. This gain came at a cost (of .09) in memory quantity, $F(1, 56) = 39.19$, $p < .001$, $\eta^2_p = .41$, which did

not change over time either, $F(2, 56) = 1.03$, $p = .364$, $\eta^2_p = .04$. Thus, as shown in Figure 5 (panel A), the quantity-accuracy tradeoff was stable over the 48-hour retention period.

Report criterion. To test our hypothesis that the decline in free-report accuracy over time could stem from setting an overly liberal report criterion at delayed testing, we estimated each participant's report criterion using a computational procedure developed by Koriat and Goldsmith (1996). Considering each confidence level (between 1% and 100%) as a candidate P_{rc} (report criterion probability), hits were defined as volunteered answers for which P_a (assessed probability) $\geq P_{rc}$, and correct rejections as withheld answers for which $P_a < P_{rc}$. The chosen P_{rc} estimate for each participant was the value that maximized the percentage of hits and correct rejections combined (*fit rate*: averaging 93% across participants). This is the confidence level for each participant above which most of the participant's answers were volunteered, and below which most were withheld. For example, consider two participants with the same set of responses and associated confidence judgments, but different volunteering decisions. Whereas P1 volunteered all answers with confidence above 50%, P2 volunteered only answers with confidence above 79%. Thus, the two participants differ in their report criterion, with P1 employing a more liberal control criterion of 51 compared to P2's more conservative report criterion of 80. As shown in Figure 6 (panel A), the report criterion averaged 59, 58, and 49 at immediate, 24-hour, and 48-hour testing, respectively, with a non-significant decline over time, $F(2, 57) = .99$, $p = .378$, $\eta^2_p = .03$.

< Insert Figure 6 about here >

In summary, what were the major findings of Experiment 1? First, memory retention, as evaluated in terms of forced-report proportion correct, was higher at immediate testing than at delayed testing, exhibiting the typical decline over time, with a relatively large initial decline and a smaller non-significant decline thereafter. Second, a parallel pattern of decline over time was found in the confidence judgments assigned to the responses, largely confined

to confidence in correct responses. Third, a substantial decrease over time was found in both free-report quantity and free-report accuracy, under conditions of high accuracy motivation induced by an explicit payoff schedule for providing correct information in the free-report phase. Was this a result of reduced monitoring effectiveness at delayed testing, changes in the control policy over time (i.e., setting a lower, more liberal report criterion), or both? The results seem to implicate monitoring effectiveness; whereas confidence in genuine memories declined over time, confidence in erroneous memories did not, such that the participants' ability to discriminate between true and false memories by using their subjective confidence judgments deteriorated with time. False memory persistence was evident in the participants' volunteering decisions such that the number of correct volunteered answers declined over time whereas the number of incorrect volunteered answers did not, resulting in a reduction in free-report accuracy over time. Finally, the stability over time of the report criterion suggests that the reduction in free-report accuracy did not ensue from changes in the control policy (hypothesis 2) but from poorer monitoring effectiveness at delayed testing (hypothesis 1).

Experiment 2

Experiment 2 was identical to Experiment 1 in all respects except one: Accurate responding was encouraged via instructions rather than via monetary payoffs, aiming for higher ecological validity and generalization. The main goal of Experiment 2 was to examine the extent to which the results of Experiment 1 would be replicated when using instructions that emphasized the importance of accurate responding, in a way that more closely resembles real-life eyewitness situations.

Method

Participants and Design. Sixty native Hebrew-speaking students (25% males and 75% females, mean age= 24.50) from the University of Haifa took part in the experiment. They were randomly and equally assigned to the immediate, 24-hour, or 48-hour retention

interval groups, with 20 participants in each group. The participants were paid 30 NIS (approximately 8 US$) for their participation in each experimental session (one session in the case of the immediate-testing group, and two sessions in the case of the delayed-testing groups).

Materials and procedure. The materials and procedure were the same as in Experiment 1 except that accurate responding in the free-report phase was induced by using instructions instead of a payoff matrix. Each participant was asked to imagine that she was the only person who had viewed the target event, and that some of the details were needed for a police investigation of a crime that had happened. She was told that if she reported incorrect information, the investigation might fail, and she was therefore requested to be as accurate as she could, and to refrain from providing low-confidence answers to avoid errors. See Appendix C (in the supplementary material) for the instructions that were presented to the participants at the various stages of Experiment 2.

Results and Discussion

As in Experiment 1, two independent judges determined for each response that was provided whether or not it was correct. The classifications made by these two judges were identical in 99.33% of the cases. A third judge determined the scoring of the controversial .67% of the responses.

Forced-report quantity. As shown in Figure 1 (panel B), forced-report proportion correct declined across the three retention intervals, $F(2, 57) = 11.71$, $p < .001$, $\eta^2_p = .29$. As in Experiment 1, pairwise comparisons revealed a significant decline (of .17) over the first 24 hours, $t(38) = 3.37$, $p = .002$, $d = 1.09$, but a non-significant decline (of .05) over the next 24 hours, $t(38) = 1.07$, $p = .290$, $d = .35$.

Confidence. As also shown in Figure 1 (panel B), confidence too declined over time, $F(2, 57) = 13.44$, $p < .001$, $\eta^2_p = .32$, with significant declines over both the first 24 hours

(.11), $t(38) = 2.08$, $p = .044$, $d = 1.37$, and the next 24 hours (.15), $t(38) = 3.22$, $p = .003$, $d = 1.05$.

Volunteering rate. A significant decline across the three retention intervals was also found for the volunteering rate (see Figure 1, panel B), $F(2, 57) = 6.11$, $p = .004$, $\eta^2_p = .18$. However, only the decline (of .15) between immediate and 24-hour testing was significant, $t(38) = 2.46$, $p = .019$, $d = .80$, whereas the smaller decline (of .07) between 24-hour and 48-hour testing was not, $t(38) = 1.01$, $p = .321$, $d = .33$. Again, due to a somewhat different pattern of volunteering rates than that of the confidence judgments, we conducted an additional ANOVA. In this analysis, the difference between confidence and volunteering rate was treated as a repeated factor, for the assessment of a possible interaction between this factor and retention interval. However, the interaction was not statistically significant, $F(2, 57) = 3.02$, $p = .056$, $\eta^2_p = .10$. Over the three retention intervals, the confidence judgments were highly diagnostic of the volunteering of the answers, averaging .65 for volunteered answers and .17 for withheld answers, $F(1, 49) = 260.68$, $p < .001$, $\eta^2_p = .84$. The relationship between confidence and volunteering remained stable over time, with a non-significant interaction between retention interval and volunteering, $F(2, 49) = 2.03$, $p = .143$, $\eta^2_p = .076$.

Calibration bias. The participants were generally well-calibrated, with a non-significant overall calibration bias score of .02, $F(1, 57) = .85$, $p = .359$, $\eta^2_p = .02$, and a non-significant interaction between calibration bias and retention interval, $F(2, 57) = 3.00$, $p = .058$, $\eta^2_p = .10$.

The adjusted normalized discrimination index (ANDI). As shown in Figure 2 (panel B), the participants' ability to differentiate correct from incorrect information, as indexed by ANDI, deteriorated across the three retention intervals, $F(1, 57) = 3.83$, $p = .027$, $\eta^2_p = .12$. However, neither the decline (of .15) over the first 24 hours nor the decline (of .09)

over the next 24 hours was significant, $t(38) = 1.58$, $p = .122$, $d = .51$, and $t(38) = 1.28$, $p = .208$, $d = .42$, respectively[3].

Discrimination index. Overall, the confidence judgments were diagnostic of the correctness of the answers, averaging .66 for correct answers and .35 for incorrect answers (discrimination index = .31), $F(1, 57) = 154.58$, $p < .001$, $\eta^2_p = .73$. However, the discrimination index also declined across the three retention intervals, with a significant interaction between correctness and retention interval, $F(2, 57) = 3.46$, $p = .038$, $\eta^2_p = .11$[4]. As for ANDI, the intermediate declines across the first and the second 24-hour intervals were not significant, $F(1, 38) = 1.33$, $p = .256$, $\eta^2_p = .03$, and $F(1, 38) = 2.26$, $p = .141$, $\eta^2_p = .056$, respectively. Thus, both measures revealed a deterioration of monitoring resolution only between immediate and 48-hour testing. Here too, the correlation between the discrimination index and ANDI was strong (.57, $p < .001$), as one would expect from two measures of the same metacognitive component (i.e., monitoring resolution).

As shown in Figure 3 (panel B), the mean confidence for correct answers declined across the entire retention period, $F(2, 57) = 10.44$, $p < .001$, $\eta^2_p = .27$. Whereas the initial small decline (of .08) over the first 24 hours was not significant, $t(38) = 1.33$, $p = .193$, $d = .43$, the larger decline (of .20) over the next 24 hours was significant, $t(38) = 3.01$, $p = .004$, $d = 1.00$. As also shown in Figure 3 (panel B), in contrast to the results in Experiment 1, the mean confidence for incorrect answers declined across the three retention intervals as well, $F(2, 57) = 3.15$, $p = .051$, $\eta^2_p = .10$, averaging .39, .39, and .27, at immediate, 24-hour, and 48-hour testing, respectively. Thus, although there was no decline whatsoever over the first

[3] ANDI was significantly different from zero at immediate testing, $t(19) = 5.40$, $p < .001$, $d = 2.48$, at 24-hour testing, $t(19) = 5.65$, $p < .001$, $d = 2.59$, and at 48-hour testing, $t(19) = 3.57$, $p = .002$, $d = 1.64$.

[4] Despite this decline in the discrimination index, the confidence judgments discriminated between correct and incorrect answers at immediate testing (discrimination index = .39), $t(19) = 7.49$, $p < .001$, $d = 3.44$, at 24-hour testing (discrimination index = .31), $t(19) = 6.76$, $p < .001$, $d = 3.10$, and at 48-hour testing (discrimination index = .23), $t(19) = 8.26$, $p < .001$, $d = 3.79$.

24 hours, $t(38) = .01$, $p = .992$, $d = .003$, there was a significant decline (of .12) in confidence for incorrect answers over the next 24 hours, $t(38) = 2.55$, $p = .015$, $d = .83$. Nonetheless, as in Experiment 1, the decline in confidence associated with correct answers was larger than the decline in confidence associated with incorrect answers (as confirmed by the significant interaction between correctness and retention interval reported above), resulting in a decline over time in the discrimination index.

Number of volunteered correct vs. incorrect answers. The differential effect of time was even more pronounced in terms of the volunteering rates of correct versus incorrect answers. Again, although more correct than incorrect answers were freely reported overall, $F(1, 57) = 23.30$, $p < .001$, $\eta^2_p = .29$, the discrimination between correct and incorrect answers in terms of the number of responses the participants decided to volunteer deteriorated across the three retention intervals, $F(2, 57) = 4.72$, $p = .013$, $\eta^2_p = .14$. However, only the decline in the first 24 hours was significant, $F(1, 38) = 4.17$, $p = .048$, $\eta^2_p = .10$, whereas the decline over the next 24 hours was not, $F(1, 38) = .66$, $p = .423$, $\eta^2_p = .02$. Interestingly, after 48 hours, the number of correct (3.25) and incorrect (2.65) volunteered answers was comparable, $t(19) = 1.06$, $p = .301$, $d = .34$. As shown in Figure 4 (panel B), this was a result of stability over time in the number of incorrect freely reported answers (averaging 2.45, 2.60, and 2.65, at immediate, 24-hour, and 48-hour testing, respectively), $F(2, 57) = .09$, $p = .918$, $\eta^2_p = .003$, in tandem with a decline over time in the number of correct volunteered answers (averaging 5.65, 3.95, and 3.25 at immediate, 24-hour, and 48-hour testing, respectively), $F(2, 57) = 9.08$, $p < .001$, $\eta^2_p = .24$. Whereas the decline in the number of correct volunteered answers over the first 24 hours was significant, $t(38) = 2.79$, $p = .008$, $d = .90$, the decline over the next 24 hours was not, $t(38) = 1.19$, $p = .240$, $d = .39$.

To what extent did the false memory persistence in terms of the volunteering rate of incorrect answers, in tandem with the decline in the volunteering rate of correct answers, influence free-report performance?

Free-report accuracy. As shown in Figure 5 (panel B), in contrast to Experiment 1, in Experiment 2 no significant decline was found for free-report accuracy across the three retention intervals (averaging .69 at immediate testing, .61 after 24 hours, and .56 after 48 hours), $F(2, 57) = 1.85$, $p = .167$, $\eta^2_p = .06$. However, a significant decline (of .13) in free-report accuracy was found between immediate and 48-hour testing, $t(38) = 2.04$, $p = .049$, $d = .66$.

Free-report quantity. As can be seen in Figure 5 (panel B), there was a substantial decline in free-report quantity over time, $F(2, 57) = 9.08$, $p < .001$, $\eta^2_p = .24$, with a significant initial decline (of .17), $t(38) = 2.79$, $p = .008$, $d = .90$, and a non-significant decline (of .07) between 24-hour and 48-hour testing, $t(38) = 1.19$, $p = .240$, $d = .39$.

Quantity-accuracy tradeoff. Comparing free-report to forced-report performance, the option of free-report allowed the participants to achieve a comparable and significant gain (of .12) in memory accuracy for the three retention intervals, $F(1, 57) = 28.50$, $p < .001$, $\eta^2_p = .33$, with a non-significant interaction between report option and retention interval, $F(2, 57) = 1.61$, $p = .210$, $\eta^2_p = .05$. This gain in accuracy came at a cost (of .07) in quantity, $F(1, 57) = 39.62$, $p < .001$, $\eta^2_p = .41$, which did not change over time, $F(2, 57) = .20$, $p = .818$, $\eta^2_p = .01$. Thus, again, a stable quantity-accuracy tradeoff was obtained over time (see Figure 5, panel B).

Report criterion. Finally, we examined the effect of time on the control policy. As in Experiment 1, the computational procedure developed by Koriat and Goldsmith (1996) for estimation of each participant's report criterion yielded a fit rate of 93% across participants. However, in contrast to Experiment 1, the decline in the report criterion across the three

retention intervals was significant here, $F(2, 57) = 4.17$, $p = .020$, $\eta^2_p = .13$. As shown in Figure 6 (panel B), although the report criterion remained constant (49) over the first 24 hours, $t(38) = .01$, $p = .990$, $d = .004$, it declined (to 32) over the next 24 hours, $t(38) = 2.72$, $p = .010$, $d = .88$.

In summary, the major findings of Experiment 1 were largely replicated in Experiment 2. Again, the predicted decline over time was found in terms of forced-report proportion correct and the confidence judgments assigned to the responses. Here too, despite the high incentive for accurate responding, free-report accuracy declined over the 48-hour retention interval. As in Experiment 1, this decline seems to ensue from reduced monitoring effectiveness: The decline over time in the participants' ability to distinguish between correct and incorrect information that came to mind apparently affected their free-report decisions in terms of the volunteering rates for correct versus incorrect answers. Thus, whereas the number of correct volunteered answers declined over the 48-hour retention interval, the number of incorrect volunteered answers remained stable. This, in turn, resulted in a larger proportion of errors among the volunteered responses, and consequently, in lower free-report memory accuracy following increasing delays.

The declines over time in both volunteering rate and report criterion were more pronounced in Experiment 2 than in Experiment 1, and reached statistical significance only in Experiment 2. One should note, though, that, as in Experiment 1, the effect of retention interval on the volunteering rate was not significantly weaker than its effect on confidence, in contrast to what one would expect if a more liberal criterion was employed over time.

General Discussion

In the present study, we examined the cognitive and metacognitive underpinnings of free-report memory accuracy over time, using the methodology developed by Koriat and Goldsmith (1996). As predicted, we found a decline in free-report accuracy, such that the

eyewitness accounts became less reliable over a 48-hour retention interval. That is, the participants failed to maintain stable accuracy rates: (a) even under conditions of free-report and high motivation for accuracy (whether induced by using an explicit payoff matrix or instructions), and (b) despite the documented ability of rememberers to strategically regulate the accuracy of their memory reports under such conditons using post-retrieval metacognitive processes of monitoring and control (e.g., Koriat & Goldsmith, 1994, 1996). As we discuss below, reduced monitoring resolution appears to be the source of the decline in free-report accuracy over time.

The metacognitive determinants of declining memory accuracy over time

Based on various observations in the literature, we derived two tentative hypotheses as to the metacognitive changes that may occur over time and account for the decline in free-report accuracy. Our first hypothesis was based on the well-established finding of *false memory persistence*, that erroneous memories tend to exhibit a milder rate of decline over time than genuine memories, under both free-report and forced-report conditions (e.g., Seamon et al., 2002; Thapar & McDermott, 2001). Confidence in such false memories has also been shown to be more stable over time than confidence in true memories (e.g., Garcia-Bajos & Migueles, 2003; Toglia et al., 1999). Thus, over time, the confidence judgments might become less diagnostic of the correctness of the candidate answers. Given the strong dependence of volunteering decisions on the monitoring output (i.e., the confidence judgments) shown in previous studies with healthy young adults (e.g., Koriat & Goldsmith, 1996; Kelley & Sahakyan, 2003; Pansky, et al., 2009; Rhodes & Kelley, 2005), as well as in the present study, the declining diagnosticity of the confidence ratings was expected to affect the volunteering decisions. Consequently, less high-confidence correct responses were expected to persist over time than high-confidence errors, resulting in a larger proportion of errors among the volunteered responses and lower output-bound accuracy over time. The

results confirmed this hypothesis: The decline over time in the confidence associated with correct information was more pronounced than the decline in the confidence associated with incorrect information, replicating previous findings obtained with different materials (e.g., Garcia-Bajos & Migueles, 2003; Toglia et al., 1999; Weinstein et al., 2010). Thus, the metacognitive ability to discriminate between correct and incorrect answers by using the subjective confidence judgments deteriorated over time. These findings are consistent with recent findings obtained by Carneiro, Garcia-Marques, Lapa, and Fernandez (2017), suggesting that the "editing out" of thematic intrusions, when testing conditions presumably allow for monitoring to occur, is more frequent at immediate than at delayed testing. We further found pronounced false memory persistence in terms of the participants' volunteering decisions. In fact, 48 hours after witnessing the target event, the number of incorrect event details the participants volunteered was comparable to the number of correct details they volunteered, such that accuracy was reduced to about 50%! These findings are consistent with earlier findings showing a larger decline over time in freely-recalled studied words than in freely-recalled false instrusions over two-day (Thapar & McDermott, 2001), one-week (Thapar & McDermott, 2001), and two-week (Seamon et al., 2002) retention intervals.

These results can be accounted by fuzzy-trace theory (e.g., Brainerd & Reyna, 1993, 2002), according to which each event item is encoded in memory at various levels of precision, from verbatim traces representing detailed episodic information to gist traces capturing its meaning. Over time, verbatim traces become inaccessible more rapidly than gist traces (e.g., Brainerd & Reyna, 1993, 1998, 2002; Dorfman & Mandler, 1994; Pansky, 2012; Pansky & Nemets, 2012; Pansky & Koriat, 2004). Accordingly, one could expect the quantity of correct verbatim-based answers (and hence the confidence in them and the tendency to volunteer them) to decline over time, but the quantity of incorrect gist-based answers (and

hence the confidence in them and the tendency to volunteer them) to remain more stable. Indeed, these are the results that we found in the present study.

A possible alternative explanation of the observed pattern of results is that the larger decline over time in confidence associated with true memories compared to that associated with false memories was due to the higher starting point of the former (see Figure 3). We see this explanation as an implausible account of our findings for two reasons. First, mean confidence for incorrect answers was substantially higher than zero at immediate testing (.40 in Exp. 1 and .35 in Exp. 2), leaving ample room for a decline over time (which was non-existent in Exp. 1 and relatively small in Exp. 2), Second, the same pattern was previously found even when the confidence in true and false memories at immediate testing was nearly identical (e.g., Toglia et al., 1999).

Our second hypothesis was that a decline in memory accuracy would stem from setting a more liberal report criterion at delayed testing in the face of reduced memory quantity. This hypothesis was based on the assumption that respondents might be reluctant to say "I don't know" or withhold their answers too often due to personal-social expectations for informativeness and cooperation (e.g., Ackerman & Goldsmith, 2008; Kelley & Sahakyan, 2003; Pansky et al., 2009). However, we found a significant lowering of the report criterion over time only in Experiment 2, and not in Experiment 1, whereas a decline in memory accuracy over time was found in both experiments (and was even more substantial numerically in Experiment 1). Second, the effect of retention interval on the volunteering rate was not weaker than its effect on the confidence judgments in either experiment, as one would expect if our second hypothesis were true. Therefore, we conclude that this hypothesis is not supported by the present findings. Of course, this conclusion does not preclude the possibility that a substantial lowering of the report criterion over time might occur under different circumstances, such as following longer retention intervals, yielding a reduction in

memory accuracy. What our findings show is that the decline in monitoring effectiveness can account for a decline in the accuracy of eyewitness reports over time without an additional lowering of the report criterion.

Interestingly, the participants in Experiments 1 and 2 were generally well calibrated, at both immediate and delayed testing. Supporting the assumption of Koriat and Goldsmith's (1996) model that confidence is strongly determined by the amount of accessible information in memory, the decline over time in the participants' confidence in the correctness of their answers paralleled the decline in the actual correctness of these answers.

Finally, we found similar effects of retention interval on free-report accuracy and the mediating memory and metamemory components for the two modes of encouraging accuracy (i.e., via an explicit payoff matrix in Experiment 1 and via instructions in Experiment 2). This similar pattern is important in demonstrating the applicability of Koriat and Goldsmith's (1996) framework and of similar frameworks (e.g., Higham, 2007) on the strategic regulation of memory accuracy to more naturalistic settings.

The accuracy of eyewitness memory reports over time

The present finding of declining accuracy of eyewitness accounts over time is consistent with those of several previous studies (e.g., Bahrick et al., 1993; Bergman & Roediger, 1999; Koriat et al., 2001; Larsson et al., 2003). However, we should note that more stable accuracy over time has been found when rememberers were allowed control over the grain size (i.e., level of precision or coarseness) of the information they reported. For example, Goldsmith et al. (2005) tested memory for quantitative information contained in a fictitious eyewitness transcript either immediately, after a day, or after a week. The participants were found to provide more coarse-grained answers with delay, thereby achieving a shallower (yet still substantial) decline in accuracy over time than the decline that would occur without the use of grain control. One should note, though, that in that study, the

participants were given control over grain size only and not report option, so they could not avoid reporting answers altogether, even if they knew they were wrong. Allowing participants control both over what information to report, and whether to report it at a fine grain size or coarse grain size, Pansky and Nemets (2012) found stable accuracy rates over a retention interval of 48 hours. Finally, Evans and Fisher (2011) questioned participants about details from a mock crime video using one of three questioning formats—free narrative, specific questioning (cued recall), or yes-no recognition—after either ten minutes or one week and found a significant decrease in the amount of correct information and in the precision of the information that was reported at delayed compared to immediate testing. However, there was only a negligible (marginally significant) decline in accuracy over this same time period. In free-narrative format, rememberers are allowed full control over which information to report and at which grain size to report it. Therefore, it may not be surprising that the accuracy of free-narratives is usually stable over time (e.g., Ebbesen & Rienick, 1998; Evans & Fisher, 2011), except, perhaps, when the original information is especially incoherent or ambiguous (e.g., Bergman & Roediger, 1999).

In the present study, control of grain size was not allowed as we tried to simulate a situation in which fine-grained information is required. Under such conditions, our results show that, in addition to the well-documented decline in memory quantity, even cooperative eyewitnesses who are highly motivated to be accurate and are allowed to decide which event information to report and which to withhold, fail to maintain stable memory accuracy at delayed testing. This decline ensued primarily from reduced metacognitive monitoring effectiveness by which the subjective confidence of eyewitnesses became less diagnostic of the correctness of the information that came to mind. In turn, this reduced monitoring effectiveness resulted in a larger proportion of errors among the freely-reported information, and, ultimately, in lower accuracy of the memory reports at delayed testing.

The decline in monitoring effectiveness over time that we found in the present study is seemingly at odds with earlier findings showing an improvement in monitoring effectiveness with delay (e.g., Nelson & Dunlosky, 1991; Rhodes & Tauber, 2011; Thiede & Dunlosky, 1994). One obvious difference between our study and these previous studies is that we used a retrospective metacognitive judgment of performance (i.e., confidence), whereas they used prospective judgments of learning. Several studies have shown that these two types of judgments rely to some extent on qualitatively different information and that they differ in the accuracy in which they predict (or assess) task performance (e.g., Dougherty, Scheck, Nelson, & Narens, 2005; see also Siedlecka, Paulewicz, & Wierzchon, 2016). More critically, perhaps, the delays which have been found to yield improvements in monitoring effectiveness are brief delays of seconds to minutes (see Rhodes & Tauber, 2011, for a review), whereas the delays for which we found a decline in monitoring effectiveness are extended delays of days. Supporting the importance of the length of the delay is the finding that the improvement in monitoring effectiveness with delay (e.g., Nelson & Dunlosky, 1991) is eliminated when the delay until the collection of JOLs is extended to one week (Baker & Dunlosky, 2006; Roebers, von der Linden, Schneider, & Howie, 2007).

Constraints on Generality

Before concluding, we wish to address the generalizability and limitations of the present results (see Simons, Shoda, & Lindsay, 2017). First, as our primary interest was to examine both the cognitive and metacognitive components posited to underlie free-report quantity and accuracy performance over time, we selected peripheral event details that were more likely to decay over time than central details (Flowe et al., 2016). Thus, although the target items in the present study were each clearly visible in the slides viewed by the participants, they were peripheral with regard to the main themes of the narrated slide show. Based on previous research showing not only differential forced-report memory performance

for central versus peripheral details (e.g., Christianson & Loftus, 1987; Flowe et al., 2016), but also differential monitoring effectiveness (e.g., Roberts & Higham, 2002), any findings obtained for central details could differ from our findings. However, as noted by Read and Connolly (2007), this distinction does not speak to the forensic importance of the details, as some peripheral details could have little to do with the event's interpretation and meaning but may be critical with regard to the guilt or innocence of a suspect.

Second, we should note that other factors that might moderate forgetting, such as enhanced emotion (e.g., Burke, Heuer, & Reisberg, 1992), which we did not investigate in the present study, may also alter the pattern of results. Third, our study was conducted with young healthy adults. Different patterns of findings could be expected for special populations that have been shown to differ from young adults with regard to both memory and metamemory performance, such as children (e.g., Koriat et al., 2001; Roebers et al., 2001; Roebers & Schneider, 2005), older adults (e.g., Kelley & Sahakyan, 2003; Pansky et al., 2009; Rhodes & Kelley, 2005), and schizophrenic patients (e.g., Danion, Gokalsing, Robert, Massin-Krauss, & Bacon, 2001; Koren, Seidman, Goldsmith, & Harvey, 2006). Finally, as a limited number of items, over a limited range of retention intervals, were examined in our study, our conclusions should be re-examined using a wider range of materials and retention intervals. From a purely theoretical perspective, given the role of gist-based false-memory persistence in our interpretation of the findings, it would be interesting for future research to examine whether the decline in monitoring effectiveness over time we obtained would disappear for (less ecological) gist-free events.

To conclude, the present study joins several recent studies similarly aiming to isolate the contributions of retrieval, metacognitive monitoring, and metacognitive control, to free-report memory performance of eyewitnesses under various conditions (e.g., McCallum et al., 2016; Portnoy & Pansky, 2016; Rechdan, Hope, Sauer, Sauerland, Ost, & Merkelbach, 2018;

Rechdan, Sauer, Hope, Sauerland, Ost, & Merkelbach, 2017; Sauer & Hope, 2016; Zawadzka, Krogulska, Button, Higham, & Hanczakowski, 2016). Our results highlight the important role of post-retrieval metacognitive monitoring in influencing free-report memory accuracy in general, and, more specficially, in influencing the faithfulness in which past events are recollected over time.

References

Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting–with and without satisficing knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*, 1224-1245.

Allwood, C. M., Innes-Ker, Å. H., Homgren, J., & Fredin, G. (2008). Children's and adults' realism in their event-recall confidence in responses to free recall and focused questions. *Psychology, Crime & Law*, *14*, 529-547.

Bahrick, H. P., Hall, L. K., & Dunlosky, J. (1993). Reconstructive processing of memory content for high versus low test scores and grades. *Applied Cognitive Psychology*, 7, 1-10.

Baker, J. M. C., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin & Review*, *13*, 60–65.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 309-338). Hillsdale, NJ: Erlbaum.

Benjamin , A. S., & Diaz , M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 73-94). NY: Psychology Press.

Bergman, E. T., & Roediger, H. L. (1999). Can Bartlett's repeated reproduction experiments be replicated? *Memory & Cognition*, *27*, 937-947.

Brainerd, C. J., & Reyna, V. F. (1993). Memory independence and memory interference in cognitive development. *Psychological Review*, *100*, 42-67.

Brainerd, C. J., & Reyna, V. F. (1998). Fuzzy-trace theory and children's false memories. *Journal of Experimental Child Psychology*, *71*, 81-129.

Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, *11*, 164-169.

Brainerd, C. J., Reyna, V. F., & Brandse, E. (1995). Are children's false memories more persistent than their true memories? *Psychological Science*, *6,* 359-364.

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, *13*, 207-230.

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

Carneiro, P., Garcia-Marques, L., Lapa, A., & Fernandez, A. (2017). Explaining the persistence of false memories: a proposal based on associative activation and thematic extraction. *Memory*, *25*, 986-998.

Christianson, S. Å., & Loftus, E. F. (1987). Memory for traumatic events. *Applied Cognitive Psychology*, *1*, 225-239.

Danion, J. M., Gokalsing, E., Robert, P., Massin-Krauss, M., & Bacon, E. (2001). Defective relationship between subjective experience and behavior in schizophrenia. *American Journal of Psychiatry*, *158*, 2064-2066.

Dorfman, J., & Mandler, G. (1994). Implicit and explicit forgetting: When is gist remembered? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *47*, 651-672.

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition*, *33*, 1096–1115.

Ebbesen, E. B., & Rienick, C. B. (1998). Retention interval and eyewitness memory for events and personal identifying attributes. *Journal of Applied Psychology*, *83*, 745-762.

Ebbinghaus, H. E. (1895). *Memory: A contribution to experimental psychology*. New-York, Dover (Republished 1964).

Evans, J. R., & Fisher, R. P. (2011). Eyewitness memory: Balancing the accuracy, precision and quantity of information through metacognitive monitoring and control. *Applied Cognitive Psychology*, *25*, 501-508.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Flowe, H. D., Takarangi, M. K., Humphries, J. E., & Wright, D. S. (2016). Alcohol and remembering a hypothetical sexual assault: Can people who were under the influence of alcohol during the event provide accurate testimony? *Memory*, *24*, 1042-1061.

Garcia-Bajos, E., & Migueles, M. (2003). False memories for script actions in a mugging account. *European Journal of Cognitive Psychology*, *15*, 195-208.

Goldsmith, M. (2016). Metacognitive quality-control processes in memory retrieval and reporting. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 357-385). NY: Oxford University Press.

Goldsmith, M., & Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. S. Benjamin & B. Ross (Eds.), *Psychology of learning and motivation, Vol. 48: Memory use as skilled cognition* (pp. 1-60). San Diego, CA: Elsevier.

Goldsmith, M., Koriat, A., & Pansky, A. (2005). Strategic regulation of grain size in memory reporting over time. *Journal of Memory and Language*, *52*, 505-525.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732-764.

Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specifity principle revisited. *Memory & Cognition*, *30*, 67-80.

Higham, P. A. (2007). No special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General, 136,* 1-22.

Higham, P. A., & Tam, H. (2005). Generation failure: Estimating metacognition in cued-recall. *Journal of Memory and Language*, *52*, 595-617.

Hollins, T. J., & Weber, N. (2017). Evidence of a metacognitive benefit to memory? *Memory*, *25*, 317-325.

Hope, L., Gabbert, F., & Fisher, R. P. (2011). From laboratory to the street: Capturing witness memory using the Self-Administered Interview. *Legal and Criminological Psychology*, *16*, 211-226.

Johnson. M.K., Hashtroudi, S., & Lindsay, D.S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3-28.

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring and control in the attainment of memory accuracy. *Journal of Memory and Language*, *48*, 704-721.

Koren, D., Seidman, L. J., Goldsmith, M., & Harvey, P. D. (2006). Real-world cognitive—and metacognitive—dysfunction in schizophrenia: a new approach for measuring (and remediating) more "right stuff". *Schizophrenia Bulletin*, *32*, 310-326.

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General*, *123*, 297-316.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory performance. *Psychological Review, 103*, 490-517.

Koriat, A., Goldsmith, M., & Pansky A. (2000). Toward psychology of memory accuracy. *Annual Review of Psychology, 51*, 481-537.

Koriat, A., Goldsmith, M., Schneider, W., & Nakash-Dura, M. (2001). The credibility of children's testimony: Can children control the accuracy of their memory reports? *Journal of Experimental Child Psychology*, *79*, 405-437.

Koriat, A., Levy-Sadot, R., Edry, E., & de Marcas, S. (2003). What do we know about what we cannot remember? Accessing the semantic attributes of words that cannot be recalled. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1095-1105.

Larsson, A. S., Granhag, P. A., & Spjut, E. (2003). Children's recall and the cognitive interview: Do the positive effects hold over time? *Applied Cognitive Psychology*, *17*, 203-214.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of subjective probabilities: The state of the art up to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). New York: Cambridge University Press.

Loftus, E. F. (2003). Make-believe memories. *American Psychologist*, *58*, 864-873.

Masson, M. E., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: Implication for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *35*, 509-527.

McCallum, N. A., Brewer, N., & Weber, N. (2016). Memorial monitoring and control: How confidence and social and financial consequences affect eyewitnesses' reporting of fine-grain information. *Applied Cognitive Psychology*, *30*, 375-386.

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language*, *35*, 212-230.

Neisser, U. (1996). Remembering as doing. *Behavioral and Brain Sciences*, *19*, 203-204.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*, 267-271.

Pansky, A. (2012). Inoculation against forgetting: Advantages of immediate versus delayed initial testing due to superior verbatim accessibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1792-1800.

Pansky, A., & Goldsmith, M. (2014). Metacognitive effects of initial question difficulty on subsequent memory performance. *Psychonomic Bulletin & Review*, *21*, 1255-1262.

Pansky, A., Goldsmith, M., Koriat, A., & Pearlman-Avnion, S. (2009). Memory accuracy in old age: Cognitive, metacognitive, and neurocognitive determinants. *European Journal of Cognitive Psychology*, *21*, 303-329.

Pansky, A., & Koriat, A. (2004). The basic-level convergence effect in memory distortions. *Psychological Science*, *15*, 52-59.

Pansky, A., Koriat, A., & Goldsmith, M. (2005). Eyewitness recall and testimony. In K. D. Williams & N. Brewer (Eds.), *Psychology and Law: An empirical perspective* (pp. 93-150). New York: Guilford.

Pansky, A., & Nemets, E. (2012). Enhancing the quantity and accuracy of eyewitness memory via initial memory testing. *Journal of Applied Research in Memory and Cognition*, *1*, 2-11.

Pansky, A., & Tenenboim, E. (2011). Interactions between spontaneous instantiation to the basic level and post event suggestions. *Memory*, *19*, 901-915.

Pansky, A., Tenenboim, E., & Bar, S. K. (2011). The misinformation effect revisited: Interactions between spontaneous memory processes and misleading suggestions. *Journal of Memory and Language*, *64*, 270-287.

Payne, D. G., Elie, C. J., Blackwell, J. M., & Neuschatz, J. S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language*, *35*, 261-285.

Penrod, S., Loftus, E. F., & Winkler, J. (1982). The reliability of eyewitness testimony: A psychological perspective. In N. L. Kerr & R. M. Bray (Eds.), *The psychology of the courtroom* (pp. 119-168). New York: Academic Press.

Portnoy, S., & Pansky, A. (2016). Metacognitive effects of initial question difficulty on subsequent eyewitness memory performance. *Journal of Applied Research in Memory and Cognition, 5*, 159-167.

Read, J. D., & Connolly, D. A. (2007). The effects of delay on long-term memory for witnessed event. In M. D. Toglia, J. D. Read, D. F. Ross, & R. C. L., Lindsay (Eds.), *The Handbook of eyewitness psychology*: *Vol. I: Memory for events* (pp. 117-155). Mahwah, NJ: Erlbaum Associates Inc.

Rechdan, J., Hope, L., Sauer, J. D., Sauerland, M., Ost, J., & Merckelbach, H. (2018). The effects of co-witness discussion on confidence and precision in eyewitness memory reports. *Memory*, *26*, 904-912.

Rechdan, J., Sauer, J. D., Hope, L., Sauerland, M., Ost, J., & Merckelbach, H. (2017). Computer mediated social comparative feedback does not affect metacognitive regulation of memory reports. *Frontiers in Psychology*, *8*, 1433.

Rhodes, M. G., & Kelly, C. M. (2005). Executive processes, memory accuracy and memory monitoring: An aging and individual differences analysis. *Journal of Memory and Language*, *52*, 578-594.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, *137*, 131-148.

Roberts, W. T., & Higham, P. A. (2002). Selecting accurate statements from the cognitive interview using confidence ratings. *Journal of Experimental Psychology: Applied*, *8*, 33-43.

Roebers, C. M., Moga, N., & Schneider, W. (2001). The role of accuracy motivation on children's and adults' event recall. *Journal of Experimental Child Psychology, 78,* 313-329.

Roebers, C. M., & Schneider, W. (2005). The strategic regulation of children's memory performance and suggestibility. *Journal of Experimental Child Psychology, 91*, 24-44.

Roebers, C. M., von der Linden, N., Schneider, W., & Howie, P. (2007). Children's metamemorial judgments in an event recall task. *Journal of Experimental Child Psychology*, *97*, 117-137.

Ronis, D. L. & Yates, J. F. (1987). Components of probability judgement accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, *40*, 193-218.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*, 734-760.

Sauer, J., & Hope, L. (2016). The effects of divided attention at study and reporting procedure on regulation and monitoring for episodic recall. *Acta psychologica*, *169*, 143-156.

Schacter, D. L., Guerin, S. A., & St. Jacques, P. L. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences*, 15, 467–474.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*, 33-45.

Seamon, J. G., Luo, C.R., Kopecky, J. J., Price, C. A., Rothschild, L., Fung, S., & Schwartz, M. A. (2002). Are false memories more difficult to forget than accurate memories? *Memory & Cognition*, *30*, 1057-1064.

Siedlecka, M., Paulewicz, B., & Wierzchoń, M. (2016). But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. *Frontiers in psychology*, *7*, 218.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123-1128.

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences*, *21*, 971-986.

Thapar, A., & McDermott, K. B. (2001). False recall and false recognition induced by presentation of associated words: Effects of retention interval and level of processing. *Memory & Cognition*, *29*, 424-432.

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology*, *86*, 290-302.

Toglia, M. P., Neuschatz, J. S., & Goodwin, K. A. (1999). Recall accuracy and illusory memories: When more is less. *Memory*, *7*, 233-256.

Tuckey, M. R., & Brewer, N. (2003). The influence of schemas, stimulus ambiguity and interview schedule on eyewitness memory over-time. *Journal of Experimental Psychology: Applied, 9*, 101-118.

Weinstein, Y., McDermott, K. B., & Chan, J. C. K. (2010). True and false memories in the DRM paradigm on a forced choice test. *Memory*, *18*, 375-384.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235-269.

Yaniv, I., Yates, J. F., & Smith, J. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, *110*, 611-617.

Zawadzka, K., Krogulska, A., Button, R., Higham, P. A., & Hanczakowski, M. (2016). Memory, metamemory, and social cues: Between conformity and resistance. *Journal of Experimental Psychology: General*, *145*, 181-199.

Figure Captions

Figure 1: Mean forced-report quantity, confidence, and volunteering rate as a function of retention interval (manipulated between subjects) in Experiment 1 (panel A) and in Experiment 2 (panel B). Error bars indicate ±1 SEM.

Figure 2: Mean ANDI (Adjusted Normalized Discrimination Index) as a function of retention interval (manipulated between subjects) in Experiment 1 (panel A) and in Experiment 2 (panel B). Error bars indicate ±1 SEM.

Figure 3: Mean confidence in correct and incorrect answers as a function of retention interval (manipulated between subjects) in Experiment 1 (panel A) and in Experiment 2 (panel B). Error bars indicate ±1 SEM.

Figure 4: Mean number of volunteered correct and incorrect answers as a function of retention interval (manipulated between subjects) in Experiment 1 (panel A) and in Experiment 2 (panel B). Error bars indicate ±1 SEM.

Figure 5: Mean forced-report quantity (QTY) and accuracy (ACC), free-report quantity (QTY), and free-report accuracy (ACC) as a function of retention interval (manipulated between subjects) in Experiment 1 (panel A) and in Experiment 2 (panel B). Error bars indicate ±1 SEM.

Figure 6: Mean report criterion as a function of retention interval (manipulated between subjects) in Experiment 1 (panel A) and in Experiment 2 (panel B). Error bars indicate ±1 SEM.