Journal of Applied Research in Memory and Cognition 1 (2012) 2–10

Contents lists available at SciVerse ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Enhancing the quantity and accuracy of eyewitness memory via initial memory testing☆

Ainat Pansky*, Elvira Nemets

Department of Psychology, University of Haifa, Haifa, Israel

ARTICLE INFO

Article history:
Received 11 May 2011
Received in revised form 22 June 2011
Accepted 30 June 2011
Available online 4 August 2011

Keywords:
Eyewitness memory
Testing effect
Recall
Memory accuracy
Confidence

ABSTRACT

Previous studies have demonstrated the benefits of initial memory testing in terms of "inoculating" eyewitness memory against forgetting. The aim of the present study was to determine to what extent and under which conditions such testing may also enhance the accuracy of subsequent retrieval. Two aspects of interpolated testing were manipulated: the mode of interpolated testing (forced verbatim vs. free level) and its timing (immediate vs. delayed). After witnessing a target event, participants were questioned about event details either immediately or after a 48-h delay, and were either required to respond at the verbatim level or were given control over the grain size of their responses. Verbatim memory for event details was finally tested 72 h after the event under both standard forced-report conditions and free-report conditions. Immediate interpolated testing was found to improve both memory quantity and memory accuracy on the final test, whereas delayed interpolated testing improved only memory quantity (and to a lesser extent). Although the mode of interpolated testing affected performance on the initial test, it had no effect on either memory quantity or memory accuracy on the final test. Practical implications with regard to eyewitness questioning are discussed.

1. Introduction

Information obtained from eyewitnesses is critical in the criminal justice process. Nevertheless, it is often incomplete (e.g., Kebbell & Milne, 1998) and sometimes even erroneous (e.g., Wells et al., 2000). Eyewitness interviews may vary along several dimensions, such as when they are conducted, whether the questions are open-ended or closed, and the level of detail that is required. The primary interest of the present study is to explore the conditions under which an initial interview may enhance later memory performance in terms of both completeness and accuracy.

1.1. Inoculation against forgetting

An impressive body of research has shown that testing must be considered not solely in terms of its evaluative potential but also in terms of its mnemonic benefits: Numerous laboratory studies have demonstrated that taking a test on studied material improves retention of that material on a subsequent memory test (the *testing effect*; for a recent review, see Roediger & Butler, 2011). Several of these studies have examined the effects of testing in simulated eyewitness situations, demonstrating benefits of intervening memory tests on subsequent memory performance (e.g., Chan & Langley, 2011; Pansky & Tenenboim, 2011; Poole & White, 1991), even when the initial test is self-administered (Gabbert, Hope, & Fisher, 2009).

These findings are consistent with the suggestion that memory testing soon after an event may have beneficial effects of "inoculating" eyewitness memory against forgetting of event details (e.g., Brainerd & Ornstein, 1991; Gabbert et al., 2009; Goodman & Quas, 2008). According to fuzzy-trace theory (FTT), an event detail is encoded in memory at various levels of precision, from verbatim traces of a target's surface form to gist traces representing a target's semantic, relational, and elaborative characteristics (Reyna & Brainerd, 1995). Over time, verbatim traces become inaccessible more rapidly than gist traces (e.g., Brainerd & Reyna, 1998). An interpolated test presented soon after an event is therefore more likely to preserve verbatim traces than a delayed test and promote superior performance on the final test that requires verbatim memory (e.g., Reyna & Titcomb, 1997). Several previous studies that examined memory for complex materials (e.g., prose passages and witnessed event) have found this expected trend of a larger benefit of earlier compared to later interpolated testing for correct recall on a delayed final test (Bergman & Roediger, 1999; Pansky, submitted for publication; Spitzer, 1939).

☆ This article is based on the master's thesis submitted by the second author to the Department of Psychology at the University of Haifa. The research was supported by European Commission Grant FP6-NEST: EYEWITMEM; 43460.

* Corresponding author at: Department of Psychology, University of Haifa, Haifa 31905, Israel. Tel.: +972 4 8249437; fax: +972 4 8249431.
E-mail address: pansky@research.haifa.ac.il (A. Pansky).

2211-3681/$ – see front matter © 2011 Society for Applied Research in Memory and Cognition. Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.jarmac.2011.06.001

1.2. Control of report option and grain size in memory reporting

In the majority of studies that implemented interpolated testing, the rememberer was required to answer all the test items exactly as they were presented for study. However, in real-life situations, eyewitnesses generally have two types of control: (1) *report option* – the option to volunteer or withhold specific items of information (i.e., to respond "don't know"; Koriat & Goldsmith, 1994, 1996b) and (2) control over the *grain size* (i.e., the level of generality or detail) of the information that is reported (see Goldsmith & Koriat, 1999, 2008). In such free-report situations, rememberers may abstain from reporting aspects of the event they are not confident about or adopt a level of generality at which they are not likely to be wrong (e.g., Goldsmith & Koriat, 1999). Indeed, several studies have shown that rememberers are able to utilize the option of free report (e.g., Koriat & Goldsmith, 1994, 1996b) and control of grain size (e.g., Goldsmith, Koriat, & Pansky, 2005; Goldsmith, Koriat, & Weinberg-Eliezer, 2002; Weber & Brewer, 2008) to increase their accuracy substantially, often with only a negligible reduction in the quantity of correctly reported information.

Several recent findings suggest that the chosen grain size of responding is affected by the decline in verbatim memory over time. 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. As the retention interval increased, the participants were found to coarsen the grain size of their answers, presumably in an attempt to maintain accuracy (see also Pansky, submitted for publication; Pansky & Koriat, 2004).

Consider a witness who has the option to choose the grain size of her answers at initial testing as well as the option to refrain from answering certain questions altogether. What impact will such unrestricted responding have on her subsequent recollection, compared to restricted testing requiring verbatim precision on each and every answer, and compared to no interpolated testing at all? We propose that the answer to this question depends on two critical factors: (1) the timing of interpolated testing and (2) how subsequent memory performance is assessed.

1.3. Memory quantity vs. memory accuracy

Koriat and Goldsmith (1996a) draw a distinction between two properties by which memory performance may be evaluated: quantity and accuracy. *Quantity* is *input-bound*, assessing the likelihood of correctly remembering an input item, whereas *accuracy* is *output-bound*, reflecting the dependability of the reported information (i.e., the likelihood that a reported item is correct). In their integrative theoretical model of free-report monitoring and control, Koriat and Goldsmith (1996b) explain how people enhance their memory accuracy when given control over memory reporting. When attempting to recount past events, people use a monitoring mechanism to assess the subjective likelihood that each item of information that comes to mind is correct. They then apply a control mechanism in order to decide whether to volunteer the best available candidate answer or not. The setting of the control threshold is assumed to depend on the relative utility of providing complete vs. accurate information: The stronger the accuracy incentive, the more selective people will be in their reporting, resulting in a quantity–accuracy tradeoff. Several studies have confirmed this prediction: Raising the accuracy incentive results in fewer volunteered answers, but a higher percentage of them are correct (e.g., Koriat & Goldsmith, 1994, 1996b).

To assess the cognitive and metacognitive components underlying free-report memory performance, Koriat and Goldsmith (1996b) have developed an experimental methodology, known as the quantity–accuracy profile (QAP) methodology. This methodology is based on a two-phase, forced-free memory test, allowing the separate examination of accuracy and quantity performance that can be achieved under various conditions. In the forced-reporting phase, the participant is required to provide a response for each test item. This phase provides information about memory retention or retrieval, which is affected as little as possible by monitoring and control processes. Following the forced-reporting phase, the participants provide confidence judgments estimating the likelihood that each response is correct, allowing the assessment of monitoring effectiveness—the extent to which the subjective confidence judgments successfully differentiate correct from incorrect answers. In the free-reporting phase, the participant is asked whether she would like to volunteer the reported response or to withhold it, with monetary incentives offered for correct volunteered responses and penalties incurred for incorrect volunteered responses. This allows the assessment of the actual levels of quantity and accuracy performance that are achieved under free-report conditions.

In an eyewitness testimony situation, both quantity and accuracy are critical. Thus, in contrast to previous studies on the testing effect, in the present study we examined the effect of interpolated testing not only under the standard forced-report conditions but also under free-report conditions, allowing the assessment of potential testing effects in terms of both memory quantity and memory accuracy, using the QAP methodology. We expected that these two measures would be differentially affected by the mode of interpolated testing: Compared to forced-verbatim testing, free-level testing was expected to be less beneficial for subsequent memory performance in terms of quantity, but more beneficial in terms of accuracy.

1.4. The present study

In the present study, the participants initially viewed a narrated slide sequence containing several target details (adapted from Pansky, Tenenboim, & Bar, 2011). After completing a short distractor task, the participants performed either an immediate or a delayed (after 48 h) interpolated memory test containing questions on half of the target items and no questions on the other half (the untested condition). Following Brainerd and Reyna (1998), gist and verbatim recall were implemented using two different hierarchical levels: basic level (e.g., CHAIR) and subordinate level (e.g., WOODEN CHAIR; see Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). The participants assigned to the free-level group were allowed to choose the grain size of their memory reporting by responding at either the verbatim or the gist level, and were also given the option to refrain altogether from answering a particular question. In contrast, the participants assigned to the forced-verbatim group were required to provide a response at the verbatim level for every question. The effects of the timing and testing mode of the interpolated test were examined on a subsequent memory test that took place 72 h after exposure to the event. In this stage, all of the participants were questioned about all the target items to a verbatim specificity. This final memory test included both a forced-reporting phase and a free-reporting phase.

We predicted that the testing mode of the interpolated test would have differential effects on memory quantity and memory accuracy, particularly when interpolated testing was delayed. First, as predicted by FTT (e.g., Reyna & Titcomb, 1997) and replicating previous findings on the effect of the timing of interpolated testing (e.g., Bergman & Roediger, 1999; Pansky, submitted for publication; Spitzer, 1939), we predicted that immediate interpolated testing would yield a larger testing effect than delayed interpolated testing, in terms of larger gain in memory quantity on the final memory test compared to the untested items. Second, following Glover (1989) who found that more complete retrieval on intervening tests yielded larger benefits for later retrieval, we predicted that forced-verbatim interpolated testing would produce larger gains in memory quantity on the final memory test than would free-level interpolated testing. This benefit of forced-verbatim testing was expected to be particularly pronounced following delayed interpolated testing, assuming that, after a delay, some verbatim information would still be accessible, yet the free-level participants would be more likely to exercise the option of withholding or coarsening the grain size of their responses (e.g., Goldsmith et al., 2005).

Compared to free-level interpolated testing, forced-verbatim interpolated testing was predicted to be less beneficial in terms of memory accuracy due to an expected enhancement in the accessibility and the confidence associated with responses that the participants would be forced to retrieve under this condition, most of which would be expected to be incorrect. Previous studies have shown that forced fabrication of event details caused participants to remember the fabricated details as actual event details at a later time (e.g., Ackil & Zaragoza, 1998; Chrobak & Zaragoza, 2008; Pickel, 2004). In a similar vein, inducing participants to report misinformation was found to increase erroneous recall on a later test and the tendency to “remember” this false information from the event (Roediger, Jacoby, & McDermott, 1996). Other studies have shown that forced-report interpolated questioning can inflate subjective confidence regarding the responses to the same questions at a later time, whether correct or incorrect (e.g., Shaw, 1996; Shaw & McClure, 1996). Assuming that most of the responses that the participants would be forced to provide on the forced-verbatim test would be incorrect, particularly after a delay, the forced-verbatim participants were expected to exhibit more over-confidence on the final memory test than the participants in the free-level group.

As pointed out by Goldsmith and Koriat (1999), people control their memory reporting according to their subjective confidence in an answer, even when the latter is not diagnostic of its correctness. Hence, the increase in accessibility and confidence induced by forced-verbatim interpolated testing, particularly with regard to incorrect answers, was expected to lead to a disproportionally large tendency to volunteer these answers on the free-report phase of the final test. Consequently, forced-verbatim interpolated testing was expected to be less beneficial than free-verbatim interpolated testing, and perhaps even detrimental, in terms of memory accuracy.

2. Method

2.1. Participants

One hundred and twenty Hebrew-speaking undergraduates from the University of Haifa participated in the experiment. They were randomly assigned to one of the four groups (immediate forced-verbatim testing, immediate free-level testing, delayed forced-verbatim testing, and delayed free-level testing), with 30 participants in each group.

2.2. Materials

A 6.5-min slide show was used as the target event. The slide show consisted of still pictures accompanied by a matching soundtrack about a day in a female student’s life. Sixteen concrete details (e.g., WOODEN CHAIR, MUSHROOM PIZZA), each presented visually on a separate slide, constituted the target items (see Appendix A, column 2).

The rest of the experiment was run using a computer program developed with E-Prime experiment-generating software. The interpolated test consisted of cued-recall questions, each referring to one of the eight target items assigned to the testing condition (see Appendix A, column 3). The questions were presented in the order in which the target items had appeared in the slide show. The participants were required to answer each interpolated question in either the free-level mode or the forced-verbatim mode, depending on the experimental group to which they were assigned. In the free-level testing group, the participants were allowed to answer each question at any level they chose (i.e., verbatim or gist) or to refrain from answering. In the forced-verbatim testing group, for each target item, the same question that was presented in the free-level mode was followed by a second question, designed to solicit a response at a more detailed (subordinate) level. Thus, after answering an initial question about an item that had appeared in the slide show (e.g., “What was in the oven . . .?”), the participant was asked a second question about that item, with her response to the first question (e.g., PIZZA) inserted in the second question (e.g., “What kind of pizza?”). The remaining eight target items were not tested in this stage of the experiment. Half of the participants in each experimental group were tested on the odd-numbered items whereas half were tested on the even-numbered items.

The final cued-recall test was identical for all the participants and included questions about all 16 target items (see column 3 in Appendix A), in the order in which the target items had appeared in the slide show. All the participants were asked to recall each target item to a verbatim specificity, even if they had to guess, using the same two-question procedure as on the interpolated forced-verbatim test.

2.3. Procedure

In the first stage of the experiment, the participants viewed the slide show. Next, they completed a non-verbal filler task of solving Raven’s Progressive Matrices for approximately 10 min. Half of the participants proceeded immediately to the second stage of the experiment in which the interpolated test was administered, whereas the other half performed this stage in a separate session, after 48 h. For each of the eight tested items, the participants assigned to the forced-verbatim testing condition were required to provide a verbatim-level response, whereas those in the free-level testing condition were given the option of free-report and control over grain size. The third and final stage was administered 72 h after the first experimental session for all the participants, and combined both forced and free reporting on an item-by item basis. After providing a response to each question, the participants were also requested to provide a confidence judgment estimating the likelihood that their response was correct, between 0 and 100%. No monetary incentive was offered for performance in the forced-reporting phase. Finally, the participants were asked to decide whether to volunteer or withhold their response (i.e., the free-reporting phase). Volunteering accurate responses was induced by a moderate-incentive payoff schedule: The participants were paid 1 NIS (.25 $US) for each correct response that was volunteered and penalized the same amount for each incorrect response that was volunteered. They were told that they would not be penalized (but neither would they receive any bonus) for withheld responses. This three-phase (i.e., forced-report response, confidence judgment, volunteer decision) procedure was repeated for each target item. See Appendix B for the instructions that were presented to the participants at the various stages of the experiment.

3. Results

Two independent judges determined for each response that was provided on the final memory test, whether or not it was correct at the verbatim level. The same judges also determined for each response provided on the free-level interpolated test whether it

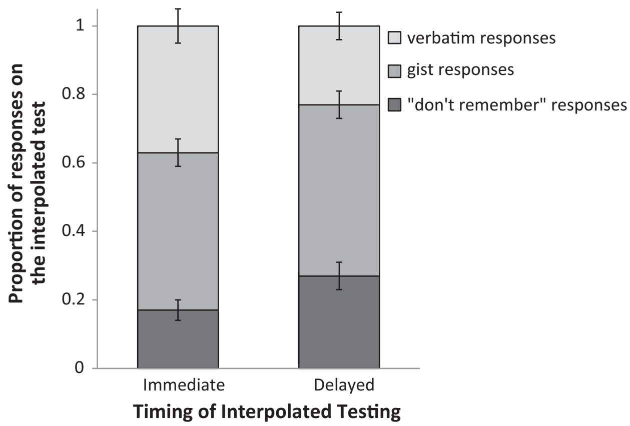

Fig. 1. Mean proportion of verbatim, gist, and "don't remember" responses provided on the free-level interpolated test at immediate vs. delayed testing (after 48 h). Error bars indicate 1 SEM.

was: (1) provided at the verbatim level, gist level, or neither ("don't remember") and (2) whether or not it was correct at the verbatim level. Finally, the judges determined for each response provided on the forced-verbatim interpolated test whether it was correct at the verbatim level. The classifications made by these two judges were identical in 97.5% of the cases (Cohen's kappa = .95). A third judge determined the scoring of the controversial 2.5% of the responses.[1]

Analyses of variance (ANOVAs) and planned comparisons (*t* tests) were used for statistical testing. All the analyses were significant at the $p < .05$ level, unless otherwise noted.

3.1. Interpolated test results

Before turning to the main analyses, we examine the data collected from the interpolated test in order to verify our assumptions regarding the influence of the mode and timing of the test.

First, the free-level responses were examined to determine whether the grain size of responding (irrespective of the correctness of the response) was affected by the timing of the test. As shown in Fig. 1, the proportion of verbatim-level responses decreased when interpolated testing was delayed (.23) compared to when it was immediate (.37), $t(58) = 2.185$, $d = 0.56$. In contrast, the proportion of gist-level responses did not vary significantly with the timing of interpolated testing (.46 and .50, for immediate and delayed testing, respectively), $t(58) = 0.578$, *ns*, $d = 0.15$. Finally, the proportion of "don't remember" responses increased when interpolated testing was delayed (.27) compared to when it was immediate (.17), $t(58) = 2.071$, $d = 0.53$. The simple interpretation of these data is that when the retention interval until the interpolated test was extended, the participants utilized the option of free-level reporting and volunteered fewer fine-grained answers, replacing them with "don't remember" responses. A second, more plausible interpretation is that there was a general shift in grain size over time: Following a delay, some answers that would have been reported at a fine grain size (i.e., verbatim) immediately were reported at a coarse grain size (i.e., gist) and some answers that would have been reported at a coarse grain size immediately were withheld entirely.

In order to directly assess verbatim memory performance, the proportion of correct responses at the verbatim level was analyzed using a two-way ANOVA, with timing of interpolated testing (immediate vs. delayed) and mode of interpolated testing (forced-verbatim vs. free-level) as between-subject factors. As expected, the timing of the interpolated test had a significant effect, $F(1, 116) = 14.173$, $\eta_p^2 = .109$, with a lower proportion of correct verbatim responses when the interpolated test was delayed (.31) than when it was immediate (.45), reflecting a decline of verbatim memory over time. In addition, the effect of interpolated testing mode was also significant, $F(1, 116) = 33.765$, $\eta_p^2 = .225$, with a higher proportion of correct verbatim responses for the forced-verbatim retrieval group (.49) than for the free-level retrieval group (.27; see Table 1, column 3). The interaction between timing and mode of interpolated testing was not significant, $F < 1$. Thus, forcing the participants to provide responses at the verbatim level yielded a larger proportion of correct verbatim responses on the interpolated test than when the level of responding was unrestricted, whether the interpolated test was immediate or delayed.

A similar analysis was performed on the incorrect verbatim responses. A significant effect of interpolated testing mode was found, $F(1, 116) = 271.058$, $\eta_p^2 = .700$, with forced-verbatim participants providing a higher proportion of incorrect verbatim responses (.51) than free-level participants (.03; see Table 1, column 4). A significant main effect of timing, $F(1, 116) = 7.201$, $\eta_p^2 = .058$, was qualified by a significant interaction between timing and mode of interpolated testing, $F(1, 116) = 7.201$, $\eta_p^2 = .058$: Whereas under free-level instructions, the same negligible proportion of incorrect responses was provided regardless of whether the interpolated test was immediate or delayed (.03), $t(58) = 0.00$, *ns*, $d = 0.00$, under forced-verbatim instructions, a higher proportion of incorrect responses was provided when testing was delayed (.58) than when it was immediate (.43), $t(58) = 2.858$, $d = 0.74$. Thus, requiring the participants to respond at the verbatim level yielded a large proportion of incorrect verbatim responses on the interpolated test, particularly after a delay.

Compared to forced-verbatim reporting, the option of free-level reporting allowed the participants to improve their memory performance in terms of the accuracy of the verbatim responses they provided by .41 (from .49 to .90), $F(1, 106) = 103.634$, $\eta_p^2 = .49$. This enhancement of accuracy was comparable at immediate (.37) and at delayed testing (.45), with a non-significant interaction between timing and mode of interpolated testing, $F(1, 106) = 1.025$, *ns*, $\eta_p^2 = .010$. By reducing the proportion of responses provided at the verbatim level between immediate and delayed testing, the free-level participants were able to maintain stable verbatim accuracy between these two testing occasions, $t(48) = 1.219$, *ns*, $d = 0.32$.

The results obtained at interpolated testing indicate the success of our manipulation of interpolated testing mode and its timing in yielding a differential pattern of responding on the interpolated test. We next examine to what extent this differential pattern of responding later affected performance on the final test.

3.2. Final test results

For each participant, six measures were calculated separately for the tested and for the untested items on the final test: (1) input-bound quantity (IBQ) – the proportion of correct verbatim responses provided in the forced-report phase out of the total number of test questions; (2) confidence – the mean confidence assigned to the responses in the forced-report phase, converted to an assessed probability of correctness ranging between 0 and 1 (by dividing each confidence rating by 100); (3) volunteer rate

[1] Responses that were provided at the correct gist level, but did not constitute a proper response at the verbatim level (e.g., DAILY NEWSPAPER instead of MA'ARIV NEWSPAPER) were omitted from the analyses. These responses, which constituted 1.4% of the total responses, were excluded because the confidence regarding their correctness reflected perceived correctness at the gist rather than at the (requested) verbatim level. That is, the confidence for these responses was inflated and would have biased the volunteer rate and the free-report IBQ and OBA measures, had these responses been included.

Table 1
Mean proportions of correct and incorrect verbatim responses on the interpolated test as a function of its timing (immediate vs. delayed) and mode of reporting (forced verbatim vs. free level). Standard deviations appear in parentheses.

Interpolated testing timing	Interpolated testing mode	Correct verbatim responses	Incorrect verbatim responses
Immediate	Free level	.34 (.25)	.03 (.08)
	Forced verbatim	.57 (.22)	.43 (.22)
Delayed	Free level	.20 (.18)	.03 (.07)
	Forced verbatim	.42 (.19)	.58 (.19)
Mean	Free level	.27 (.22)	.03 (.08)
	Forced verbatim	.49 (.22)	.51 (.22)

Table 2
Mean forced-report input-bound quantity (IBQ), confidence, volunteer rate, free-report input-bound quantity (IBQ), and output-bound accuracy (OBA) on the final memory test as a function of interpolated test timing (immediate vs. delayed), interpolated testing mode (forced verbatim vs. free level), and testing condition (untested vs. tested items). Standard deviations appear in parentheses.

Interpolated testing timing	Interpolated testing mode	Testing condition	Forced-report IBQ	Confidence	Volunteer rate	Free-report IBQ	Free-report OBA
Immediate	Free level	Untested items	.30 (.18)	.50 (.15)	.45 (.25)	.24 (.15)	.60 (.35)
		Tested items	.56 (.22)	.76 (. 13)	.72 (.22)	.51 (.25)	.68 (.26)
	Forced verbatim	Untested items	.32 (.19)	.49 (.15)	.35 (.17)	.23 (.15)	.63 (.35)
		Tested items	.63 (.22)	.76 (.17)	.71 (.24)	.57 (.25)	.79 (.18)
Delayed	Free level	Untested items	.35 (.24)	.52 (.16)	.44 (.24)	.30 (.22)	.67 (.34)
		Tested items	.40 (.21)	.59 (.17)	.52 (.24)	.33 (.20)	.60 (.28)
	Forced verbatim	Untested items	.29 (.20)	.51 (.16)	.42 (.25)	.26 (.19)	.59 (.37)
		Tested items	.41 (.18)	.63 (.16)	.59 (.21)	.36 (.18)	.60 (.27)
Mean	Free level	Untested items	.33 (.21)	.51 (.16)	.44 (.24)	.27 (.19)	.63 (.34)
		Tested items	.48 (.23)	.68 (.18)	.62 (.25)	.42 (.24)	.64 (.27)
	Forced verbatim	Untested items	.30 (.20)	.50 (.16)	.38 (.22)	.25 (.17)	.61 (.36)
		Tested items	.52 (.23)	.70 (.18)	.65 (.23)	.47 (.24)	.70 (.26)

– the proportion of volunteered responses (whether correct or incorrect) in the free-report phase out of the total number of test questions; (4) over-confidence – the difference between the participant's mean assessed probability correct and the actual proportion correct in the forced-report phase; (5) free-report IBQ – the proportion of correct verbatim responses volunteered in the free-report phase out of the total number of test questions; and (6) output-bound accuracy (OBA) – the proportion of correct verbatim responses volunteered in the free-report phase out of the total number of volunteered answers.

Each of the six dependent measures was subjected to a separate mixed-model ANOVA, with timing of interpolated testing (immediate and delayed) and interpolated testing mode (free-level and forced-verbatim) as between-subject factors, and interpolated testing condition (tested and untested) as a within-subject factor.[2]

3.2.1. *Input-bound quantity (IBQ)*

Table 2 (column 4) presents the mean forced-report IBQ on the final memory test in each of the experimental cells. A significant testing effect was found, with a higher proportion of correct verbatim responses for previously tested items (.50) than for previously untested items (.32), $F(1, 116) = 53.990$, $\eta_p^2 = .318$. Surprisingly, comparable IBQ was found following free-level and forced-verbatim interpolated testing, $F < 1$. Moreover, contrary to our predictions, a comparable testing effect was found for the two interpolated testing modes, with a non-significant interaction between interpolated testing mode and testing condition, $F(1, 116) = 1.678$, *ns*, $\eta_p^2 = .001$.

As expected, IBQ was affected by the timing of interpolated testing, with higher IBQ on the final test following immediate interpolated testing (.45) than following delayed interpolated testing (.36), $F(1, 116) = 9.694$, $\eta_p^2 = .077$. This main effect was qualified by a significant interaction between interpolated testing condition and timing of interpolated testing, $F(1, 116) = 16.642$, $\eta_p^2 = .125$. As shown in Fig. 2, the timing of interpolated testing affected only the tested items, yielding a larger testing effect following immediate (.28) than following delayed interpolated testing (.08), although both testing effects were significant, $t(59) = 8.002$, $d = 1.03$, and

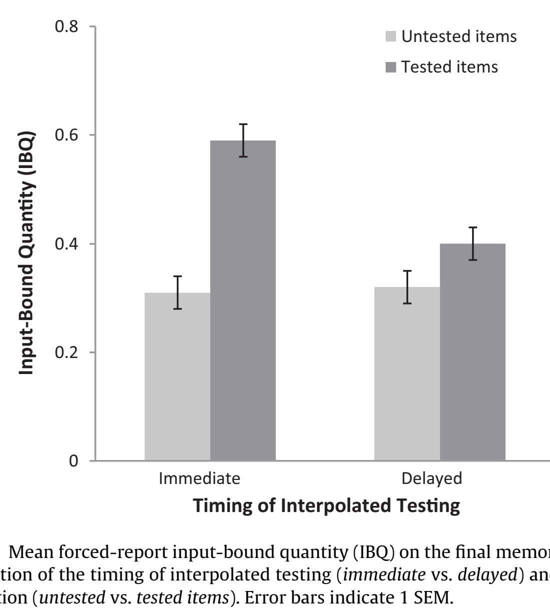

Fig. 2. Mean forced-report input-bound quantity (IBQ) on the final memory test as a function of the timing of interpolated testing (*immediate* vs. *delayed*) and testing condition (*untested* vs. *tested items*). Error bars indicate 1 SEM.

[2] All the analyses were also carried out with version of the interpolated test (whether the tested items were the odd or even target items, and vice versa for the untested items) as an additional variable. Although version sometimes interacted with the independent variables of interest, the inclusion of this variable in the analyses did not compromise any of the reported effects or conclusions. Therefore, the reported analyses were conducted with the data pooled across the two versions.

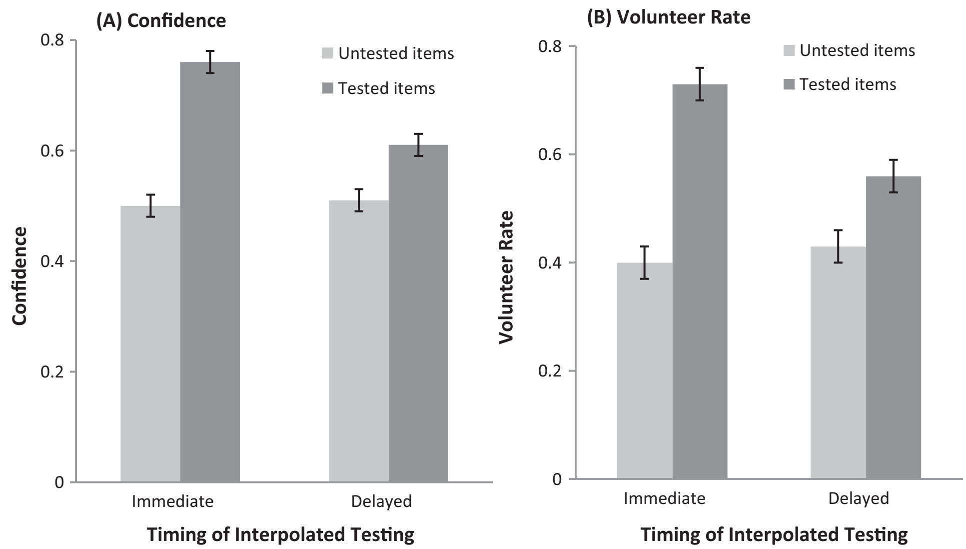

Fig. 3. Mean confidence (panel A) and volunteer rate (panel B) on the final memory test as a function of timing of interpolated testing (*immediate* vs. *delayed*) and testing condition (*untested* vs. *tested items*). Error bars indicate 1 SEM.

$t(59) = 2.339$, $d = 0.30$, respectively. Neither the interaction between timing and mode of interpolated testing, $F(1, 116) = 1.449$, *ns*, $\eta_p^2 = .012$, nor the interaction between timing of interpolated testing, its mode, and testing condition, $F < 1$, was significant.

3.2.2. *Confidence and volunteer rate*

A highly similar pattern of results was found for confidence and volunteer rate, therefore we report the findings for these two measures jointly (see Table 2, columns 5 and 6). Both confidence and volunteer rate were higher for previously tested items (.69 and .63, respectively) than for previously untested items (.51 and .41, respectively), with a significant testing effect obtained for each of these measures, $F(1, 116) = 94.492$, $\eta_p^2 = .449$, for confidence, and $F(1, 116) = 75.339$, $\eta_p^2 = .394$, for volunteer rate. Contrary to our predictions, the effect of interpolated testing mode was not significant for either of these measures, $F < 1$, nor was the interaction between interpolated testing mode and testing condition [$F < 1$, for confidence, and $F(1, 116) = 2.904$, *ns*, $\eta_p^2 = .024$, for volunteer rate].

The timing of interpolated testing influenced both confidence and volunteer rate, which were higher following immediate (.63 and .56, respectively) than following delayed interpolated testing (.56 and .49, respectively), $F(1, 116) = 9.794$, $\eta_p^2 = .078$, for confidence, and $F(1, 116) = 3.793$, $p > .055$, $\eta_p^2 = .032$ (marginally significant), for volunteer rate. As expected, a significant interaction was found between interpolated testing condition and its timing on confidence, $F(1, 116) = 21.080$, $\eta_p^2 = .154$. As shown in Fig. 3 (panel A), interpolated testing increased confidence to a greater extent when it was immediate (by .26) than when it was delayed (by .10). A parallel interaction was found for volunteer rate, $F(1, 116) = 13.935$, $\eta_p^2 = .107$: Interpolated testing increased volunteer rate to a greater extent when it was immediate (by .32) than when it was delayed (by .13; see Fig. 3, panel B). In addition, for both measures, neither the interaction between the mode of interpolated testing and its timing [$F < 1$, for confidence, and $F(1, 116) = 1.425$, *ns*, $\eta_p^2 = .012$, for volunteer rate] nor the interaction between interpolated testing mode, timing of interpolated testing, and testing condition ($F < 1$, for both measures), was significant.

To summarize, the findings for confidence and volunteer rate parallel those obtained for IBQ: All three measures were higher for previously tested than for previously untested items, with more pronounced testing effects following immediate than following delayed interpolated testing. Contrary to our predictions, interpolated testing mode was not found to influence any of these measures.

3.2.3. *Calibration (over-confidence)*

Due to the small number of observations in each experimental cell (8), the only feasible measure of monitoring effectiveness was calibration. Calibration was assessed as the degree of over-confidence: the difference between the mean assessed-probability-correct (i.e., confidence judgment) and the actual proportion of correct answers (i.e., IBQ). Overall, the participants were over-confident in the correctness of their responses (by .19), $t(119) = 13.67$, $d = 1.27$. However, the degree of over-confidence was not affected by any of the independent variables [$F < 1$, for all the effects and interactions except for the interaction between the mode of interpolated testing and its timing, $F(1, 116) = 2.270$, *ns*, $\eta_p^2 = .019$]. Thus, the parallel pattern of results obtained for IBQ and confidence resulted in a comparable degree of over-confidence in all the experimental cells.

3.2.4. *Free-report IBQ and OBA performance*

Finally, we examined free-report memory performance in terms of the quantity (i.e., IBQ) and accuracy (i.e., OBA) of the freely reported information (see Table 2, columns 7 and 8). Six participants were omitted from the following analyses because they chose not to volunteer any of their responses in the untested condition in the free-report phase of the final test.

A significant testing effect was found in terms of free-report quantity performance, with superior IBQ for the tested (.44) than for the untested items (.26), $F(1, 110) = 64.638$, $\eta_p^2 = .370$. Although OBA was somewhat higher for the tested (.67) than for the untested items (.62), the testing effect in terms of OBA was not significant, $F(1, 110) = 1.719$, *ns*, $\eta_p^2 = .015$. As for the previous measures, the effect of interpolated testing mode was not significant for either IBQ or OBA, $F < 1$, nor was the interaction between interpolated testing mode and testing condition, $F(1, 110) = 2.230$, *ns*, $\eta_p^2 = .020$, for IBQ, and $F(1, 110) = 1.476$, *ns*, $\eta_p^2 = .013$, for OBA.

IBQ was higher following immediate (.39) than following delayed interpolated testing (.31), $F(1, 110) = 6.247$, $\eta_p^2 = .054$, whereas OBA was not affected by the timing of interpolated testing, $F(1, 110) = 1.564$, *ns*, $\eta_p^2 = .014$. As expected, though, interpolated testing yielded a larger testing effect when it was immediate than when it was delayed. As shown in Fig. 4 (panel A), interpolated test-

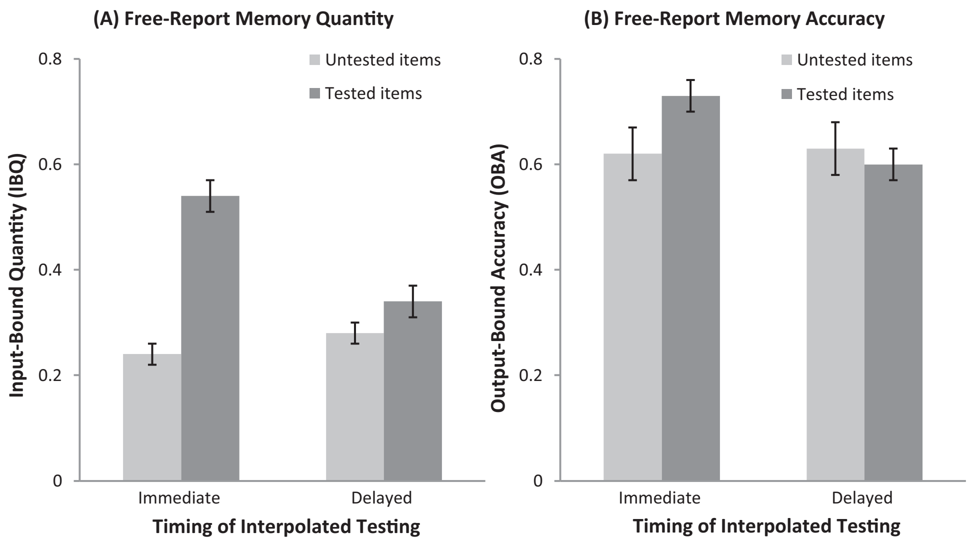

Fig. 4. Mean free-report input-bound quantity (IBQ; panel A) and output-bound accuracy (OBA; panel B) on the final memory test as a function of timing of interpolated testing (*immediate* vs. *delayed*) and testing condition (*untested* vs. *tested items*). Error bars indicate 1 SEM.

ing increased IBQ to a greater extent when it was immediate (by .30) than when it was delayed (by .06), with a significant interaction between interpolated testing condition and interpolated testing timing, $F(1, 110) = 28.490$, $\eta_p^2 = .206$. As shown in Fig. 4 (panel B), the same pattern was found for OBA, with a larger testing effect following immediate interpolated testing (.11) than following delayed interpolated testing (−.03), $F(1, 110) = 5.269$, $\eta_p^2 = .046$. In fact, only immediate interpolated testing yielded a significant testing effect in terms of OBA, $t(57) = 2.820$, $d = 0.37$, whereas delayed interpolated testing did not, $t(55) = 0.668$, *ns*, $d = 0.09$. Finally, the interaction between the mode of interpolated testing and its timing was significant neither for IBQ, $F < 1$, nor for OBA, $F(1, 110) = 1.250$, *ns*, $\eta_p^2 = .011$. Similarly, the interaction between interpolated testing mode, timing of interpolated testing, and testing condition was not significant for either of the two measures, $F < 1$.

4. Discussion

In the present study, we examined the effect of the mode of interpolated testing (forced-verbatim or free-level) and its timing (immediate or after 48 h) on subsequent recall performance under both forced-report and free-report conditions. We predicted that the traditional testing effect would be found not only in terms of memory quantity (both forced-report and free-report), but also in terms of free-report memory accuracy. We also expected, based on FTT (see Brainerd & Reyna, 1998), that this effect would be more pronounced following immediate than following delayed interpolated testing. These predictions were supported. Finally, we predicted that the mode of interpolated testing would have a differential effect on quantity and accuracy: Compared to forced-verbatim interpolated testing, free-level interpolated testing was expected to yield inferior memory quantity but superior memory accuracy on the final test. This prediction was not confirmed, as neither memory quantity nor memory accuracy was affected by interpolated testing mode.

4.1. *Effects of the timing of interpolated testing*

Replicating the findings of numerous studies that have demonstrated superior memory quantity for previously tested than for previously untested items (for a review, see Roediger & Butler, 2011), we found a testing effect in terms of forced-report memory quantity. Consistent with earlier studies (e.g., Bergman & Roediger, 1999; Pansky, submitted for publication; Spitzer, 1939), we also found that immediate interpolated testing improved the subsequent recall of previously tested items to a greater extent than delayed interpolated testing. This finding is consistent with the reduction in the proportion of responses that were freely provided at the verbatim-level and the decline in the quantity of correct verbatim responses, which we found between immediate and delayed interpolated testing (see also Goldsmith et al., 2005; Pansky, submitted for publication; Pansky & Koriat, 2004). These findings are in line with the assumption of FTT that the smaller advantage of delayed compared to immediate interpolated testing is due to the gradual decay of verbatim traces (Brainerd & Reyna, 1998; Reyna & Titcomb, 1997).

The parallel pattern of findings that was obtained for confidence (as for IBQ) is consistent with earlier findings supporting the view that confidence in an answer depends on the amount and strength of evidence retrieved in support of that answer (e.g., Griffin & Tversky, 1992; Koriat, Lichtenstein, & Fischhoff, 1980). Supporting the view that people volunteer and withhold information on the basis of subjective confidence (e.g., Goldsmith & Koriat, 1999; Koriat & Goldsmith, 1996b), we found that immediate interpolated testing yielded higher volunteer rates on the final test than delayed interpolated testing.

A similar pattern of results was found for free-report memory quantity: Whereas both immediate and delayed interpolated testing were found to yield significant testing effects, these effects were larger following immediate than following delayed interpolated testing. With regard to free-report accuracy, only immediate interpolated testing was found to improve it, whereas delayed interpolated testing did not. A close examination of other measures assessed in this condition reveals that compared to no interpolated testing, delayed interpolated testing was found to increase the proportion of volunteered items to a larger extent (by 0.13) than it increased the proportion of correct responses (by only 0.08). Thus, it seems that delayed testing induced participants to volunteer almost as many more wrong responses as more correct responses. This, in turn, resulted in no significant improvement in free-report accuracy as a result of (delayed) interpolated testing.

4.2. *Effects of interpolated testing mode*

The mode of interpolated testing did not influence either memory quantity or memory accuracy. As expected, free-level testing yielded a lower proportion of correct verbatim recall than forced-verbatim testing on the interpolated test. However, when they were required to provide responses at the verbatim level on the final test, the free-level participants performed as well as the forced-verbatim participants.

One possible explanation for the surprisingly high memory quantity of the free-level participants on the final test is that at least some of the responses provided at the gist level on the interpolated test were covertly retrieved correctly at the verbatim level. This covert retrieval may have successfully inoculated their subsequent memory against forgetting. We assume that these verbatim memories, though correct, were not reported due to a relatively strict report criterion that seems to have been adopted by the free-level participants. Note that these participants were tested under unrestricted instructions that provided no incentives for informativeness. Indeed, given the near-zero proportion of incorrect verbatim responses provided on the free-level interpolated test, it seems that the participants were highly selective in their verbatim reporting (and also successful in terms of the accuracy of the responses that they did chose to provide at the verbatim level), sacrificing informativeness for accuracy. Perhaps, in an attempt to achieve optimal accuracy on the interpolated test, the free-level participants preferred to report verbatim level responses only when those were held with very high confidence and chose to provide coarse-grained answers or even withhold their response entirely when they experienced even the slightest uncertainty.

Free-report memory accuracy performance was also comparable following free-level and forced-verbatim interpolated testing. Our findings suggest that forcing participants to report items at the verbatim level, whether correct or not, increased their average confidence in the correctness of those items when reported on a subsequent memory test to the same extent as free-level interpolated reporting wherein participants were allowed to choose which items to report at the verbatim level. Thus, consistent with some earlier findings (e.g., Henkel, 2004; Kang et al., 2011; McDermott, 2006) but inconsistent with others (e.g., Ackil & Zaragoza, 1998; Chrobak & Zaragoza, 2008; Pickel, 2004; Roediger et al., 1996), forced guessing on one occasion did not affect performance on a later occasion. It is possible that covert retrieval of both correct and incorrect answers on the free-level interpolated test not only enhanced memory quantity on the final test but also increased confidence (and, consequently, volunteer rate) to the same extent as forced-verbatim responding, resulting in comparable accuracy performance. It is also possible that merely repeating the test question via interpolated testing, regardless of whether the verbatim response was retrieved or not, increased subsequent confidence to the same extent in both conditions by increasing cue familiarity (for a similar idea, see Chandler, 1994).

4.3. Practical application

Given that memory accuracy, or the dependability of reported information, is of paramount concern in the setting of eyewitness testimony, the timing of initial eyewitness questioning should be considered seriously. The present findings highlight the importance of immediate eyewitness questioning in acquiring more dependable subsequent reports. However, both memory goals (i.e., quantity and accuracy) are important in the context of eyewitness testimony. In light of the improvement in memory quantity following delayed interpolated testing, albeit smaller than the improvement following immediate interpolated testing, it seems that conducting a delayed interpolated test is better than conducting no interpolated testing at all.

In the present study, we also manipulated the mode of initial testing as either unrestricted free-level interpolated testing or more demanding forced-verbatim interpolated testing. As Poole and White (1995) have pointed out, witnesses to crimes are typically queried many times before they are asked to testify in court. Therefore, determining which method of initial questioning – requiring verbatim precision or free-level unrestricted questioning – is more beneficial to eyewitness memory performance at a later point is important from an applied perspective. Our results show that the two modes of interpolated recall contribute equally to the completeness and reliability of the subsequent memory report. Nonetheless, it is not only memory performance on the final test that is important in a practical setting. Memory performance at earlier points in time may be of critical importance as well (e.g., for a police investigation). In determining which mode of reporting is better at the time of interpolated testing, one must decide which memory goal is more important, memory quantity or memory accuracy.

Finally, an important potential criticism regarding the applicability of the testing effect literature to real-world settings such as the education and eyewitness realms is that the documented effects of interpolated testing merely involve teaching rememberers to produce a fixed response when given a particular retrieval cue, thus drilling a particular (verbatim) response (see Roediger & Butler, 2011). The present findings of comparable benefits to verbatim memory performance on the criterial test following free-level interpolated testing (during which the participants often did not provide a verbatim response) as following forced-verbatim interpolated testing, converge with other recent findings (e.g., Butler, 2010; Pansky & Tenenboim, 2011; Rohrer, Taylor, & Sholar, 2010) in suggesting that the benefits of testing extend beyond rote repetition and retention of a specific response.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jarmac.2011.06.001.

References

Ackil, J. K., & Zaragoza, M. S. (1998). Memorial consequences of forced confabulation: Age differences in susceptibility to false memories. *Developmental Psychology*, *34*, 1358–1372.

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

Brainerd, C. J., & Ornstein, P. A. (1991). Children's memory for witnessed events: The developmental backdrop. In J. Doris (Ed.), *The suggestibility of children's recollections* (pp. 10–20). Washington, DC: APA.

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

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1118–1133.

Chan, J. C. K., & Langley, M. (2011). Paradoxical effects of testing: Retrieval enhances both accurate recall and suggestibility in eyewitnesses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 248–255.

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, *22*, 273–280.

Chrobak, Q. M., & Zaragoza, M. S. (2008). Inventing stories: Forcing witnesses to fabricate entire fictitious events leads to freely reported false memories. *Psychonomic Bulletin & Review*, *15*, 1190–1195.

Gabbert, F., Hope, L., & Fisher, R. P. (2009). Protecting eyewitness evidence: Examining the efficacy of a self-administered interview tool. *Law and Human Behavior*, *33*, 298–307.

Glover, J. A. (1989). The testing phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.

Goldsmith, M., & Koriat, A. (1999). The strategic regulation of memory reporting: Mechanisms and performance consequences. In D. Gopher, & A. Koriat (Eds.), *Attention and performance XVII—Cognitive regulation of performance: Interaction of theory and application* (pp. 373–400). Cambridge, MA: MIT Press.

Goldsmith, M., & Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. Benjamin, & B. H. Ross (Eds.), *The psychology of learning and motivation* (pp. 1–60). London: Academic Press.

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.

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size memory reporting. *Journal of Experimental Psychology: General*, *131*, 73–95.

Goodman, G. S., & Quas, J. A. (2008). Repeated interviews and children's memory: It's more than just how many. *Current Directions in Psychological Science*, *17*, 386–390.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435.

Henkel, L. A. (2004). Erroneous memories arising from repeated attempts to remember. *Journal of Memory and Language*, *50*, 26–46.

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology*, *103*, 48–59.

Kebbell, M. R., & Milne, R. (1998). Police officers' perceptions of eyewitness performance in forensic investigations. *Journal of Social Psychology*, *138*, 323–330.

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*, 297–315.

Koriat, A., & Goldsmith, M. (1996a). Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences*, *19*, 167–228.

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

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.

McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition*, *34*, 261–267.

Pansky, A. (submitted for publication). Inoculation against forgetting: Advantages of immediate vs. delayed initial testing due to superior verbatim accessibility.

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

Pansky, A., & Tenenboim, E. (2011). Inoculating against eyewitness suggestibility via interpolated verbatim vs. gist testing. *Memory & Cognition*, *39*, 155–170.

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.

Pickel, K. L. (2004). When a lie becomes the truth: The effects of self-generated misinformation on eyewitness memory. *Memory*, *12*, 14–26.

Poole, D. A., & White, L. T. (1991). Effects of question repetition on the eyewitness testimony of children and adults. *Developmental Psychology*, *27*, 975–986.

Poole, D. A., & White, L. T. (1995). Tell me again and again: Stability change in the repeated testimonies of children adults. In M. S. Zaragoza, J. R. Graham, G. C. N. Hall, R. Hirschman, & Y. S. Ben-Porath (Eds.), *Memory and testimony in the child witness* (pp. 24–43). Thousand Oaks, CA: Sage.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: Some foundational issues. *Learning and Individual Differences*, *7*, 145–162.

Reyna, V. F., & Titcomb, A. L. (1997). Constraints on the suggestibility of eyewitness testimony: A fuzzy-trace theory analysis. In D. G. Payne, & F. G. Conrad (Eds.), *A synthesis of basic and applied approaches to human memory* (pp. 157–174). Hillsdale, NJ: Erlbaum.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*, 20–27.

Roediger, H. L., Jacoby, J. D., & McDermott, K. B. (1996). Misinformation effects in recall: Creating false memories through repeated retrieval. *Journal of Memory and Language*, *35*, 300–318.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*, 233–239.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Shaw, J. S. (1996). Increases in eyewitness confidence resulting from postevent questioning. *Journal of Experimental Psychology: Applied*, *2*, 126–146.

Shaw, J. S., & McClure, K. A. (1996). Repeated postevent questioning can lead to elevated levels of eyewitness confidence. *Law and Human Behavior*, *20*, 629–653.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656.

Weber, N., & Brewer, N. (2008). Eyewitness recall: Regulation of grain size and the role of confidence. *Journal of Experimental Psychology: Applied*, *14*, 50–60.

Wells, G. L., Malpass, R. S., Lindsay, R. C. L., Fisher, R. P., Turtle, J. W., & Fulero, S. M. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist*, *55*, 581–598.