Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Memory and Language 52 (2005) 505–525
Journal of Memory and Language
www.elsevier.com/locate/jml

Strategic regulation of grain size in memory reporting over time ☆

Morris Goldsmith *, Asher Koriat, Ainat Pansky

Department of Psychology, University of Haifa, Haifa, Israel

Received 9 September 2004; revision received 2 January 2005
Available online 21 February 2005

Abstract

As time passes, people often remember the gist of an event though they cannot remember its details. Can rememberers exploit this difference by strategically regulating the "grain size" of their answers over time, to avoid reporting wrong information? A metacognitive model of the control of grain size in memory reporting was examined in two experiments, in which memory for quantitative information contained in a fictitious eyewitness transcript was tested either immediately, or after 1 day or 1 week. Given control over the grain size of their answers, participants' report accuracy still declined with delayed testing, but at a slower rate than when forced to provide only precise answers, remaining stable between the 1-day and 1-week tests. The observed pattern of change and stability in chosen-grain report accuracy was traced to the use of a stable control policy in the face of less effective monitoring judgments at delayed testing.

Keywords: Memory accuracy; Metamemory; Monitoring and control; Confidence judgments; Gist memory; Forgetting

A large amount of research has shown that people can remember the gist of an event although they cannot remember its details. Much of that research has examined gist versus verbatim memory of linguistic-textual information. The basic finding is that the general meaning of studied material is forgotten less rapidly than is more precise information, such as the surface form or verbatim wording of that material (e.g., Begg & Wickelgren, 1974; Kintsch, Welsch, Schmalhofer, & Zimny, 1990; Reyna & Kiernan, 1994). Kintsch et al. (1990), for example, found differential forgetting rates for three different levels of textual information, with surface information (verbatim memory) becoming inaccessible within four days, memory for the semantic content (gist) declining at a slower rate, and judgments based on situational memory (valid inferences from a relevant knowledge schema) remaining highly stable over time. Other findings involving similar comparisons also indicate that coarse information is forgotten less rapidly than more precise information: for example, with respect to the forgetting of the semantic attributes of studied words versus the words themselves (Koriat, Levy-Sadot, Edry, & de Marcas, 2003), of academic course concepts versus specific facts (Conway, Cohen, & Stanhope, 1991), of fictional character roles versus character names (Stanhope, Cohen, & Conway, 1993), and of thematic-superordinate versus subordinate story propositions (Christiaansen, 1980; Kintsch, Kozminsky, Streby, McKoon, & Keenan, 1975).

☆ This research was supported by a grant to Morris Goldsmith and Asher Koriat from the Israel Science Foundation founded by the Israel Academy of Sciences and Humanities. Facilities for conducting the research were provided by the Institute of Information Processing and Decision Making, University of Haifa, and by the Max Wertheimer Minerva Center for Cognitive Processes and Human Performance. We thank Amit Weinberg-Eliezer and Sarah Kate Bar for their help in preparing and running the experiments, and in reviewing the relevant literature.
* Corresponding author. Fax: +972 4 8249431.
E-mail address: mgold@research.haifa.ac.il (M. Goldsmith).

0749-596X/$ - see front matter
doi:10.1016/j.jml.2005.01.010

Different explanations of this pattern have been put forward. Most prominently, multi-representational theories (e.g., Brainerd & Reyna, 1990; Kintsch et al., 1990; Neisser, 1986; Reyna & Brainerd, 1995) posit independent representations at various (hierarchical) levels of precision with differential decay rates over time: the more precise-verbatim level traces decaying at a faster rate than the more general-gist level traces. Other, more limited explanations, based on schema abstraction and reconstructive processing (e.g., Bransford & Franks, 1971; Neisser, 1984; see Alba & Hasher, 1983; Brewer & Nakamura, 1984, for reviews) and differential attention patterns during encoding (e.g., Gernsbacher, 1985; Murphy & Shapiro, 1994) have also been proposed.

Regardless of the theoretical explanation of the phenomenon, several basic implications of this literature cast the notion of forgetting in a new light. First, memory and forgetting are clearly not all-or-none phenomena: people's memory performance varies depending on the level of detail or coarseness (i.e., the "grain size") at which it is measured. Hence, the time course of forgetting cannot be described in terms of a single "forgetting curve" (Ebbinghaus, 1895/1964). Second, with the passage of time, memory undergoes not only quantitative but also qualitative changes, such that different types (levels) of information can be accessed or reconstructed at different points in time. This implies that when people are allowed to decide for themselves what level of information to provide, for example, on the witness stand, they may have to make strategic decisions regarding the level at which to report each piece of information, and these decisions too may change over time. In this regard, it is curious that although some theories have incorporated the idea that people can control the hierarchical level on which they base their recognition memory responses (e.g., Anderson, Budiu, & Reder, 2001; Brainerd, Wright, Reyna, & Payne, 2002), there has been virtually no work on how such control might contribute to recall memory performance.[1]

[1] It is interesting to note that virtually all of the laboratory findings regarding the forgetting rates for coarse versus detailed information have been based on forced-recognition memory testing (e.g., comparisons between hit rates for studied items and false-acceptance rates for non-studied items that share semantic/categorical/situational content with the studied items at various hierarchical levels; Kintsch et al., 1990), in which the grain size of the responses is completely controlled by the experimenter. Such procedures are well suited to identify the levels of representation that remain available to the rememberer at different points in time—and perhaps to infer some ensuing constraints on potential recall performance. However, they do not allow one to trace the changes in the accuracy and graininess of actual recall performance over time, nor to specify how these changes might be mediated by the strategic regulation of memory grain size.

In fact, evidence for the idea that personal control over the level of responding might contribute to the level of achieved recall performance over time is primarily anecdotal. For example, Neisser (1988), in explaining the superior accuracy of open-ended recall testing over forced-choice recognition testing in his naturalistic study, noted that whereas the recognition format required making relatively fine discriminations, the recall format allowed participants to choose "a level of generality at which they were not mistaken" (1988; p. 553; but see also Koriat & Goldsmith, 1994). Along similar lines, Fisher (1996), in assessing participants' freely reported recollections of a filmed robbery found no difference in the accuracy of statements made soon after the event and statements made 40 days later. This seeming anomaly was resolved when the grain size of the statements was considered: the information reported at the longer interval tended to be more coarse than the information reported at the shorter interval.

The general hypothesis implied by these two examples is that in recalling episodic information from memory, people may choose to provide more coarsely grained answers as the retention interval increases, thereby maintaining a reasonably high and stable level of report accuracy over time, but at the expense of providing less precise/detailed information. This hypothesis is consistent with the findings reviewed earlier, that detailed information suffers a faster forgetting rate than coarse information, and findings from recognition-memory research, that memory responses may be strategically based on more coarse levels of representation when the detailed information becomes harder to access (Anderson et al., 2001; Brainerd et al., 2002). It also could perhaps help explain some puzzling results in the eyewitness memory literature, indicating surprisingly high and stable levels of recall accuracy over periods of up to several years (e.g., Ebbesen & Rienick, 1998; Flin, Boon, Knox, & Bull, 1992; Hudson & Fivush, 1991; Poole & White, 1991, 1993). In these naturalistic studies, like in the two preceding examples, the remembered information was elicited in an open-ended, free-narrative format, which presumably allowed the participants to choose both which information to provide, and at what level of detail or generality to report it.

Do rememberers strategically control the grain size of their answers in reporting information at different points in time? If so, what are the mechanisms by which such control is realized? What are the consequences of this control for memory accuracy and forgetting? These are the questions addressed in this article. We now turn to the theoretical framework and some working hypotheses that will guide our investigation.

A model and working hypothesis

In previous work (Goldsmith & Koriat, 1999; Goldsmith, Koriat, & Weinberg-Eliezer, 2002; Koriat & Goldsmith, 1994, 1996a, 1996b), we put forward a general theoretical framework for addressing the strategic regulation of memory reporting. The basic premise of this framework is that when recollecting information from memory, people do not simply emit all of the items of information that come to mind. Rather, they use metacognitive monitoring and control processes to decide whether to report the item at all (or else respond "don't know"; control of *report option*), and if so, at what level of precision or generality to report it (control of *grain size*). They do so first, by engaging a monitoring process which evaluates the subjective probability that each item that comes to mind is correct, and then by invoking a control mechanism which operates by way of a report criterion on the monitoring output: an item of information is reported if its assessed probability of being correct passes the criterion, otherwise, if there is a coarser grained answer whose assessed-probability-correct passes the criterion, this coarser answer will be provided (control of grain size). If not, the item will be withheld entirely (control of report option).

We propose that the dynamic that guides the control processes for both report option and grain size is an *accuracy-informativeness trade-off*: assuming that there is at least a moderate relationship between the monitoring judgments and the actual correctness of each answer, setting a higher report criterion will result in a higher proportion of correct answers out of those reported (i.e., output-bound accuracy; see Koriat & Goldsmith, 1994, 1996b). However, increases in accuracy will generally come at the cost of reduced informativeness—fewer items of information may be volunteered (report option), and those that are volunteered will tend to be less precise (grain size). Thus, the rememberer must weigh the competing incentives for accuracy and informativeness in order to arrive at the appropriate control policy (criterion level) for a particular memory situation. Empirical studies focusing on the control of report option alone (holding grain size constant; Koriat & Goldsmith, 1994, 1996b; Koriat, Goldsmith, Schneider, & Nakash-Dura, 2001) and on the control of grain size alone (holding report option constant; Goldsmith et al., 2002), have yielded good support for the model, and demonstrated its general utility (see also Danion, Gokalsing, Robert, Massin-Krauss, & Bacon, 2001; Higham, 2002; Kelley & Sahakyan, 2003; Koren et al., 2004; Payne, Jacoby, & Lambert, 2004; Roebers, Moga, & Schneider, 2001).

Can this general model help shed light on the puzzlingly flat forgetting curves observed for naturalistic recall, mentioned earlier? Suppose that memory for an experienced event is tested at different points in time using an open-ended recall procedure that allows participants the option of responding at different levels of precision or coarseness (though, because we are focusing exclusively on the control of grain size, we will not give them the option to withhold an answer entirely, i.e., report option). What should be the basic form of the "forgetting curve" in terms of changes in the accuracy and informativeness of the provided answers over time?

Fig. 1 presents three hypothetical and schematic forgetting curves, plotted in terms of the accuracy (output-bound proportion correct; A) and informativeness (average precision; B) of the provided answers on immediate testing, and two later points in time. Two of the curves are assumed to represent boundary conditions, derived from the literature reviewed earlier on differential forgetting rates for coarse (gist) and precise (verbatim) information over time (cf. Fig. 2 in Kintsch et al., 1990; Fig. 1 in Stanhope et al., 1993): in the *precise-grain* curve, the participants always provide a very precise-informative answer, regardless of the retention interval

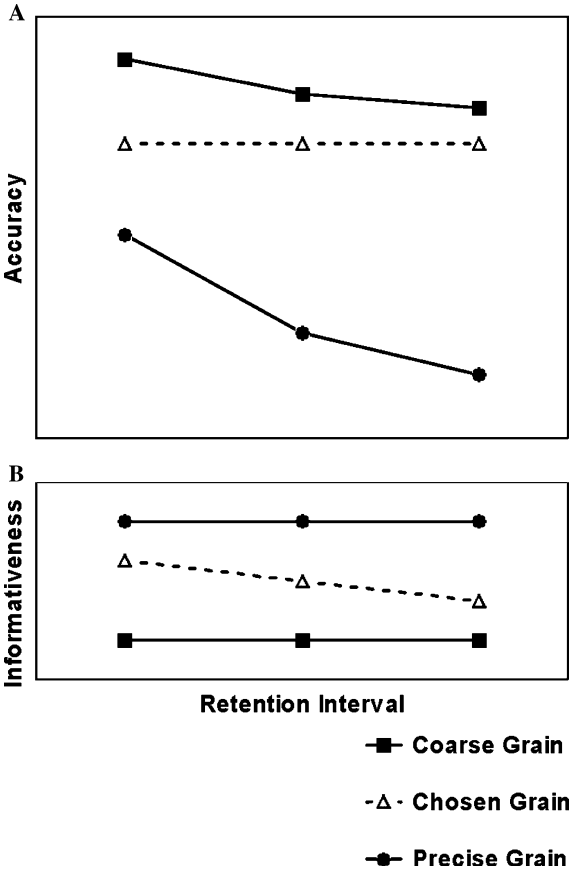

Fig. 1. Hypothetical forgetting curves depicting a plausible pattern of memory performance over time, in terms of accuracy (A) and informativeness (B), for precise-grain answers, coarse-grain answers, and "chosen-grain" answers for which the grain size is under the control of the rememberer.

(and regardless of their confidence in the answer). Accuracy in this case drops relatively steeply and then begins to level off (resembling the classic Ebbinghaus forgetting curve; see Wixted & Ebbesen, 1991, 1997), whereas the average precision of the answers is high and unchanging across the retention intervals. Conversely, in the *coarse-grain* curve, the participants always provide a very coarse-uninformative answer, regardless of the retention interval. Accuracy in this case is always higher than in the precise-only condition, and drops more slowly over time, whereas the average precision of the answers is low and unchanging across the retention intervals.

The third curve, representing a hypothetical chosen-grain condition, is the most interesting. It illustrates a plausible pattern of performance if participants regulate the grain size of their answers according to the basic model just outlined. We assume that participants are motivated to be as informative as possible, but also to provide information that has a reasonably high chance of being correct. By our model, one way of achieving this is to set a report criterion that reflects the minimum acceptable probability of being correct, and then provide the most informative answer with an assessed-probability-correct that passes this criterion. What will happen if such a control mechanism is applied to the candidate answers that are available at each point in time, given the hypothetical boundary conditions imposed by the precise and coarse forgetting curves in Fig. 1?

On immediate testing, we assume that many precise candidate answers with high (above-criterion) associated assessed probabilities will come to mind, so that a relatively high proportion of these answers will be provided. At later retention intervals, however, fewer precise answers will pass the report criterion; hence, more and more coarse-grained answers will be provided as time passes, leading to a drop over time in the average precision-informativeness of the answers that are provided (B). However, if the participants apply the same report criterion at each point in time, and their monitoring judgments (assessed probability correct) remain equally valid (or invalid), then output-bound accuracy should remain relatively stable across the retention intervals. Thus, assuming that the rememberer's monitoring and control processes are relatively stable over time (this assumption will be examined later), the metacognitive control of grain size can, in principle, yield stable report accuracy over time even though memory for both precise and coarse information drops over the same interval. This hypothetical result will be treated as a working hypothesis, guiding the two experiments reported next.

Experiment 1

Experiment 1 adapted a procedure used previously by Goldsmith et al. (2002) to examine the control of grain size in answering semantic-memory (general-knowledge) questions (see also Yaniv & Foster, 1995, 1997). Participants read a short fictitious "police interview" describing events surrounding a bar-room argument and later assault on one of the persons involved. The interview was contrived to contain various items of quantitative information: heights, weights, and ages of the characters, dates and times of day, distances, speeds, numbers of people present, and so forth. A memory questionnaire tested the participants' memory for 22 of these items either immediately, or after a 1-day or 1-week retention interval. Following Goldsmith et al. (2002), the memory questionnaire was administered in two phases: in the first phase, the participants were required to give their best answer to each item at two different grain sizes: (a) a precise value, and (b) a bounded interval of values, the width of which was specified by the experimenter. For example, "How much was the bill that Shay paid to Benny Sharone as compensation?" (A) Provide the precise amount; (B) Provide a 20-shekel (5-dollar) interval. In addition, the participants were asked to make confidence judgments for each answer, by assessing the probability (0–100%) that their answer was correct (for precise answers) or that the correct value was contained within the specified interval (for coarse answers). Finally, in the critical, second phase, the participants were asked to go over their answers again (after the confidence ratings were removed), and for each item, to indicate which of the two answers (i.e., which of the two grain sizes) they would prefer to volunteer, in order to provide the best information that they can about the facts of the case.

What are the predictions? Regarding forced-grain performance in the first test phase, accuracy (proportion correct) for both the precise and coarse answers should decrease over time, with a slower (shallower) drop for the coarse information than for the precise information. With regard to monitoring, based on previous research (e.g., Goldsmith et al., 2002), we expect a moderate relationship between confidence and the correctness of the answers at each grain size, with some degree of overconfidence. Because of the scarcity of relevant prior work on changes in monitoring effectiveness over time (see General discussion), our working hypothesis is that monitoring will be relatively stable.

Regarding the control process, the pattern of results obtained by Goldsmith et al. (2002) was seen to support a simple "satisficing" (Simon, 1956) model of the control process, in which rememberers attempt to provide the most informative (precise) answer that they can, as long as its assessed probability of being correct passes some preset criterion level. An alternative model was rejected in that study—a "relative utility" model, in which rememberers choose a grain size that maximizes the expected subjective utility of each reported answer (i.e., the assessed probability that the answer is correct multiplied by the

subjective gain obtained if it is correct, minus the assessed probability that it is wrong multiplied by the subjective penalty incurred if it is wrong). An interesting question is whether the satisficing model will apply to the control of grain size in episodic memory reporting as well. This question will be addressed by examining whether the control decision is based on confidence in the precise-grained answer alone (satisficing model) or on the relative disparity between confidence in the precise- and coarse-grained answers (relative utility model). A further question is whether the control policy will remain stable over time.

As presented earlier, with regard to the chosen-grain performance in the second test phase, our working hypothesis is that stability in the monitoring and control processes will yield relatively high levels of report accuracy that will be stable across the three retention intervals. This will be achieved by providing increasing numbers of coarse answers as retention interval increases, thereby reducing the average informativeness of the answers over time.

Method

Participants

Seventy-two Hebrew-speaking undergraduates from the University of Haifa participated in the experiment for payment (NIS 35, approximately $8). They were randomly assigned to the immediate, 1-day, or 1-week retention-interval conditions.

Materials

A 570-word text (in Hebrew) was developed, containing a fictitious police interview of witnesses to an argument that took place between two young men (Benny and Shay) in a pub, leading to a later assault on one of them (Shay). The text included 22 target items, all of which were quantitative pieces of information (e.g., the height of the assailant, the time he left the pub). A memory questionnaire was also devised, with one question on each target item, which participants were required to answer at two different grain sizes: (1) precise—a specific value, and (2) coarse—a bounded interval of a specified width. Blanks were provided next to each question for providing an answer at the two different grain sizes. For example:

How much did Shay weigh?
(A) His exact weight (in kg) ____
(B) 10 kg interval ____ - ____

For simplicity, the participants were instructed to treat the specified interval as specifying the arithmetic difference between the two endpoints of their answer (e.g., 70–80 would be considered a 10 kg interval although it is in fact an 11 kg interval).

The specified intervals for the coarse-grain alternatives differed for each item. The items and intervals were chosen on the basis of pretesting to avoid ceiling-level accuracy for the coarse-grain size on immediate testing (mean accuracy about 75%), and to avoid floor-level performance for the precise-grain size at delayed testing (mean accuracy about 25%). The order of the questions in the memory test matched the chronological order of events described in the police interview.

Procedure

The experiment was administered individually or in small groups and was divided into a study phase and two test phases. Each participant was given a separate instruction booklet for each phase and proceeded through the phase at his or her own pace. The experimenter was present at all times for clarifications.

In the study phase, participants were asked to read carefully the text describing the target event. They then performed a 5-min filler task, in which they rated their attitudes toward various aspects of the Israeli legal system. Following this (either immediately, or in a later session), the memory test was administered in two phases: in test-phase 1, the participants were given the 22-item memory questionnaire, and were required to provide the best answer that they could for each item at both grain sizes, even if they had to guess. They were also asked to make a confidence judgment regarding each answer, representing the assessed probability (0–100%) that this answer was correct (for the precise grain size) or that the correct value was contained in the specified interval (for the coarse grain size). Finally, in test-phase 2, the memory questionnaire filled out in test-phase 1 was returned to the participants, with their answers still filled in but with the confidence judgments removed (cut away). The participants were instructed to go over their answers to each question again (without changing them), and choose one answer for each item (i.e., at one of the two alternative grain sizes), according to the following rationale: "Assume that the original transcript concerning the circumstances of Shay's injury was lost. Your job is to help the investigator reproduce the facts of the case by providing one answer to each of the questions you answered earlier. For each question, please choose the answer that you prefer to provide to the investigator." The participants marked their choices by circling one of the two alternative answers for each item. They were not allowed to change any answers.

The participants assigned to the immediate-testing condition performed all three phases in a single experimental session lasting approximately 30 min. Those assigned to the other two conditions performed the study phase in the first experimental session, and returned either 1 day or 1 week later to perform the two test phases in a second experimental session.

Results and discussion

Out of 1584 observations (22 items × 72 participants), 16 observations were omitted from the analyses due to minor procedural problems such as the participant deviating from the specified grain size, omitting an answer, or using illegible handwriting. All significance tests are omnibus analyses of variance (ANOVA) or specific contrasts using the omnibus error term, unless stated otherwise.

Accuracy performance

Accuracy scores (output-bound percent correct) were calculated separately for the answers provided by each participant at the precise and coarse grain sizes in phase 1 of the memory test, and at the chosen grain size (precise or coarse) in phase 2 of the test. Fig. 2 presents the mean accuracy scores for each grain size (precise, coarse, and chosen) at the immediate, 1-day, and 1-week retention intervals, as well as the percentage of coarse answers that were chosen in phase 2 at each retention interval.

Beginning with the forced-grain answers in phase 1, the overall pattern of forgetting resembles the general pattern emerging from the forced-recognition literature discussed earlier: faster forgetting of precise than of coarse information (cf. Fig. 1, earlier). The accuracy of the precise answers showed an overall drop of 37% points over the 1-week period, $F(2,69) = 49.57$, $MSE = 181.45$, $p < .001$: from 58% on immediate testing to 30% after one day, $F(1,69) = 51.21$, $MSE = 181.45$, $p < .001$, with a further decrease to 21% after one week, $F(1,69) = 5.85$, $MSE = 181.45$, $p < .05$. The accuracy of the coarse answers dropped by 26% points over the same period, $F(2,69) = 28.07$, $MSE = 150.79$, $p < .001$: from 76% on immediate testing to 58% after one day, $F(1,69) = 24.72$, $MSE = 150.79$, $p < .001$, with a further drop to 50% after one week, $F(1,69) = 5.61$, $MSE = 150.79$, $p < .05$. Overall, the accuracy of the coarse answers (61%) was higher than for the precise answers (36%), $F(1,69) = 502.00$, $MSE = 45.26$, $p < .001$, and the forgetting rate was slower from immediate to 1-day testing, $F(1,69) = 13.80$, $MSE = 45.26$, $p < .001$, but equivalent from 1-day to 1-week testing, $F < 1$.

Turning now to the performance in phase 2 of the memory test, in which the participants were allowed to control the grain size of their answers, as expected, the participants chose to provide more coarse-grain answers as the retention interval increased, $F(1,69) = 23.97$, $MSE = 252.94$, $p < .001$: the mean percentage of coarse-grain answers increased from 43% on immediate testing to 61% after one day, $F(1,69) = 14.00$, $MSE = 252.94$, $p < .001$, with a further increase to 75% after one week, $F(1,69) = 10.08$, $MSE = 252.94$,

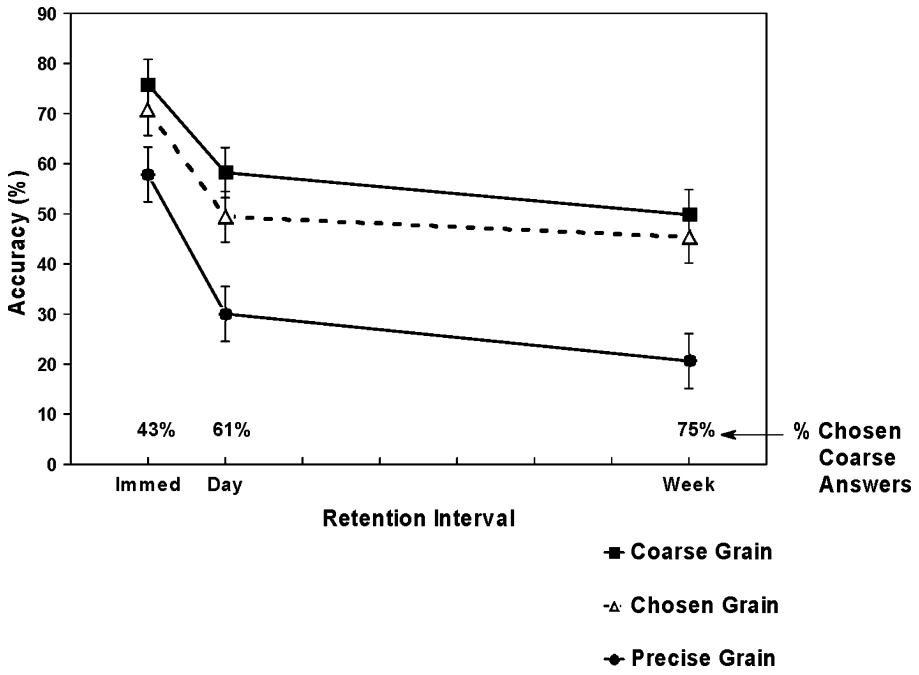

Fig. 2. Forgetting curves showing actual memory accuracy performance (mean percent correct) as a function of retention interval for the participants in Experiment 1, plotted separately for precise-grain answers (test phase 1), coarse-grain answers (test phase 1), and "chosen-grain" answers (test phase 2) for which grain size was under the control of the participant. The reduction in the informativeness of the chosen-grain answers at each retention interval (percentage of chosen coarse answers) is also separately indicated, the error bars represent 95% confidence intervals.

$p < .005$. In fact, two 1-week participants did not provide any precise-grain answers at all.

What were the consequences of this control for report accuracy? As can be seen in Fig. 2, in contrast to our working hypothesis, there was a substantial overall decline in chosen-grain memory accuracy, with a 26%-point drop in accuracy observed over the 1-week retention period, $F(2,69) = 28.99$, $MSE = 154.24$, $p < .001$. However, although the initial drop in chosen-grain accuracy from the immediate test (71%) to the 1-day test (49%) was significant, $F(1,69) = 35.26$, $MSE = 154.24$, $p < .001$, the level of chosen-grain accuracy remained stable between the 1-day test and the 1-week test (45%), $F(1,69) = 1.34$, $MSE = 154.24$, $p = .25$. Thus, although our working hypothesis was not supported in comparing immediate and 1-day (or 1-week) testing, it was supported in comparing 1-day and 1-week testing.

Moreover, in comparing chosen-grain performance with performance at the forced precise-grain size, not only was overall accuracy substantially higher for the chosen-grain answers (55% vs. 36%), $F(1,69) = 466.92$, $MSE = 27.64$, $p < .001$, but there was a noticeably smaller decrease in accuracy over time across the entire retention period, $F(1,69) = 15.14$, $MSE = 27.64$, $p < .001$, for the interaction: the precise-grain forgetting curve was more steep than the chosen-grain curve, both in the drop from the immediate to the 1-day test (by 7 percentage points), $F(1,69) = 9.28$, $MSE = 27.64$, $p < .005$, and in the drop from the 1-day to the 1-week test (by 5% points), $F(1,69) = 5.98$, $MSE = 27.64$, $p < .02$. Also, although the forgetting function for the chosen-grain answers resembles that of the forced coarse-grain answers, with the overall accuracy of the chosen-grain answers (55%) only slightly lower than the accuracy of the coarse-grain answers (61%), $F(1,69) = 55.23$, $MSE = 25.00$, $p < .001$, the shape of the forgetting curve was somewhat different, $F(2,69) = 2.55$, $MSE = 25.00$, $p = .09$, for the interaction: the curve tended to be slightly steeper for the chosen-grain answers than for the coarse-grain answers over the first day of the retention period, $F(1,69) = 3.23$, $MSE = 24.96$, $p = .08$, but was significantly flatter for the chosen-grain answers than for the coarse-grain answers over the 1-day to 1-week retention interval, $F(1,69) = 4.33$, $MSE = 24.96$, $p < .05$.

It appears, then, that there is some support for the working hypothesis that accuracy will be stable over time when participants are given control over the grain size of their answers (cf. Fig. 1). First, the accuracy of the chosen-grain answers did not decline at all between the 1-day and 1-week tests. This is in contrast to the forgetting functions for both the coarse-grain and the precise-grain answers, which exhibited significant declines over both retention intervals. Second, although there was a decline in the accuracy of the chosen-grain answers from the immediate test to the 1-day retention interval, the drop was significantly shallower than the corresponding rate of forgetting of the precise-grain answers. Thus, by controlling the grain size of their answers, the participants were able to reduce the decline in accuracy, even though they did not prevent it entirely. Of course, this relative stability in accuracy came at the price of providing answers that were informationally more coarse (i.e., of lower utility; Yaniv & Foster, 1995) as time passed.

Before finalizing this conclusion, however, we should point out an unforeseen constraint on chosen-grain accuracy that may have prevented the participants from maintaining stable accuracy over the immediate to 1-day retention interval in this experiment. Note that participants achieved a chosen-grain accuracy rate of 71% on immediate testing, but the highest accuracy that they could achieve by choosing coarse-grained answers after one day was only 58%! Apparently, in order to maintain stable accuracy, the participants would have needed to provide substantially coarser answers than were allowed to them on the 1-day test. Whether or not they would have done so, under a less restricted choice of grain size, is an open question that will be addressed later (see Experiment 2).

In sum, given the constraints imposed by the level of coarse-grain performance in this experiment, stable free-grain accuracy over time could be obtained only across the interval from 1-day to 1-week testing, and in fact it was: across this interval, there was no significant decrease in chosen-grain accuracy despite the fact that both precise-grain and coarse-grain accuracy did decrease significantly over the same interval. In fact, the pattern on this interval is quite similar to the predicted pattern (see Fig. 1).

Underlying metacognitive mechanisms

So far, we have focused on the performance consequences of the control of grain size over time. We now turn to examine the metacognitive processes underlying this control. The most basic question is whether the grain decisions were systematic. Indeed, they were: as predicted, participants tended to choose the coarse-grained answer when the more precise answer was relatively unreliable. Across the three retention intervals, answers that the participants chose to provide at the precise grain level had a relatively high (65%) chance of being correct, whereas the precise-grain answers that they chose not to provide had a relatively low (15%) chance of being correct, $F(1,67) = 316.82$, $MSE = 257.29$, $p < .001$, for the difference. This implies not only that the participants were able to distinguish between precise answers that were more likely and less likely to be correct (effective monitoring), but that they also controlled their reporting accordingly. The accuracy of the chosen precise-grain answers, however, became lower as the retention interval increased (84, 61, and 47% at the immediate, 1-day, and 1-week tests, respectively), $F(2,67) = 17.54$, $MSE = 469.38$, $p < .001$. This could stem from poorer

monitoring effectiveness over time, a lower report criterion over time, or both. Examination of each of these processes—monitoring and control—in turn, will shed light on the various possibilities.

Monitoring effectiveness. The effectiveness of memory monitoring can be evaluated in terms of *calibration* (absolute correspondence) and in terms of *resolution* (relative correspondence). Calibration refers to the overall correspondence between the assessed and actual probabilities of being correct. The calibration data from each retention interval are presented graphically in Fig. 3, plotted separately for the precise and coarse-grain answers, according to the procedure commonly used in calibration research (Lichtenstein, Fischhoff, & Phillips, 1982): the probability assessments for the answers in phase 1 were grouped into 12 levels (0.0, .01–.10, .11–.20, . . . , .91–.99, 1.0). The proportion correct is plotted against the mean assessed probability correct for the answers in each category, computed across participants. Perfect calibration is indicated by the diagonal line.

Overall there is a positive relationship between assessed and actual probability correct, though the relationship appears stronger for the precise answers than for the coarse answers, and on immediate testing (Fig. 3A) than on delayed testing (Figs. 3B and C). The general pattern of deviation from the diagonal is consistent with that of previous calibration studies (Erev, Wallsten, & Budescu, 1994), indicating overconfidence for answers with high assessed probabilities of being correct, and

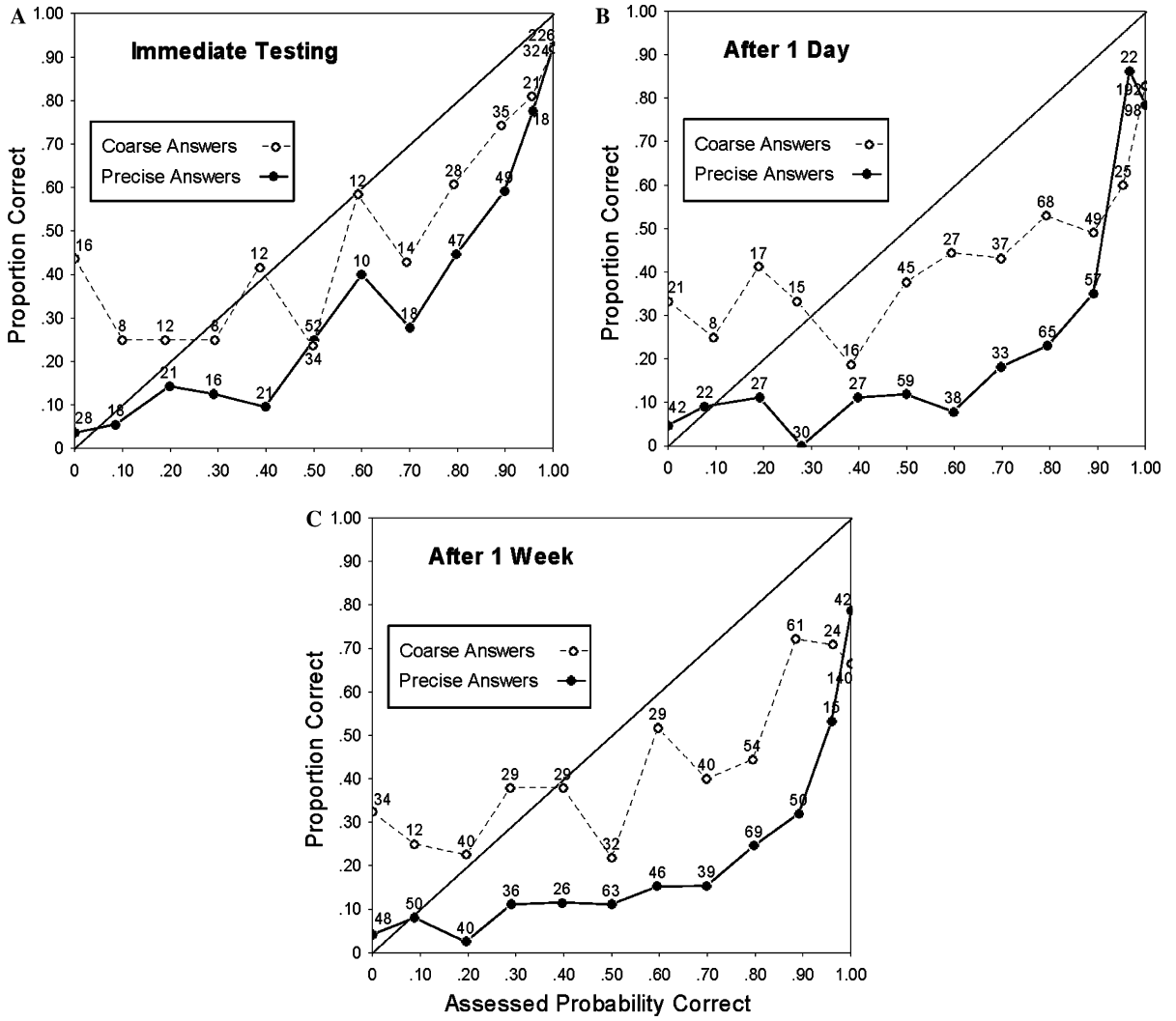

Fig. 3. Calibration curves for the precise-grain and coarse-grain answers in Experiment 1 (test phase 1), at immediate (A), 1-day (B), and 1-week (C) testing. The number of items in each category appears beside each data point.

underconfidence for answers with low assessed probabilities. This pattern is more pronounced for the coarse-grain answers than for the precise-grain answers, the latter exhibiting a large degree of overconfidence across most of the confidence range.

Our working hypothesis was that monitoring calibration would be stable over time. A comparison of the plots for immediate versus delayed testing, however, suggests otherwise: for both precise-grain and coarse-grain answers the overconfidence bias appears to increase with delayed testing. To evaluate this increase, calibration-overconfidence scores were calculated for each participant at each grain size, as the difference between the mean assessed-probability-correct and the actual proportion correct of the items answered by that participant at each grain size in phase 1 of the experiment. Confirming what can be seen visually in the plots, overall, the participants showed more overconfidence for the precise answers (.27) than for the coarse-grain answers (.14), $F(1,69) = 98.35$, $MSE = 0.0057$, $p < .001$, and overconfidence increased for delayed compared to immediate testing (.12, .25, and .24, for immediate, 1-day, and 1-week testing, respectively), $F(2,69) = 6.10$, $MSE = 0.0396$, $p < .001$. This increase was more pronounced for the precise answers (.16, .32, and .32, for immediate, 1-day, and 1-week testing, respectively) than for the coarse-grained answers (.08, .17, and .17, for immediate, 1-day, and 1-week testing, respectively), $F(2,69) = 3.52$, $MSE = 0.0057$, $p < .05$ for the interaction, but was significant in both cases.

This finding yields the potentially important implication that rememberers' monitoring judgments are not sufficiently attuned to the reduction in memory accuracy that occurs over time. Thus, even if rememberers intend to achieve the same level of accuracy at longer retention intervals as at immediate testing, they may fail to do so because of increasing overconfidence: answers that are deemed to have, say, an 80% chance of being correct after one day or after one week are less likely to be correct than answers held with the same subjective confidence on immediate testing. Calibration (miscalibration) was stable from 1-day to 1-week testing, however, which may have contributed to the stability of chosen-grain accuracy over this time segment.

Calibration is one aspect of monitoring effectiveness. A second aspect, monitoring resolution (or discrimination accuracy, Yaniv, Yates, & Smith, 1991), reflects people's ability to distinguish between correct and incorrect answers irrespective of the absolute levels of the confidence judgments. It is this aspect of monitoring that supports effective decisions about which answers can be provided at a relatively precise grain size (i.e., answers that are correct at the precise level), and which might benefit from the use of a wider interval (i.e., answers that are incorrect at the precise level). We calculated monitoring resolution for each participant at each grain size as the within-participant Kruskal–Goodman gamma correlation (see Nelson, 1984) between the assessed probability correct of each answer and whether or not the answer was correct.

Overall, monitoring resolution was higher for the precise-grain answers (.78) than for the coarse-grain answers (.55), $F(1,67) = 25.09$, $MSE = 0.0726$, $p < .001$. More importantly, resolution too decreased with delayed testing (.78, .65, and .57 for immediate, 1-day, and 1-week testing, respectively), $F(2,67) = 6.66$, $MSE = 0.0761$, $p < .005$. Like the result for calibration, monitoring resolution dropped significantly in comparing immediate and 1-day testing, $F(1,67) = 5.24$, $MSE = 0.3988$, $p < .05$, but not over the interval from one day to one week, $F(1,67) = 1.85$, $MSE = 0.3988$, $p = .18$. There was no Retention Interval × Grain Size interaction, $F < 1$.

Taken together, the initial decline and later stability in monitoring effectiveness over time—found for both resolution and calibration—mirrors the pattern of change over time in the accuracy of the chosen-grain answers in this experiment. Thus, perhaps part of this change can be explained in terms of the contribution of metacognitive monitoring to output-bound memory accuracy (for an in-depth treatment of this contribution, see Koriat & Goldsmith, 1996b).

Note, however, that the levels of monitoring effectiveness are quite good, even at delayed testing. For example, interpreted in probabilistic terms, a gamma correlation of .5 implies that if a participant were presented with two answers, one correct and one incorrect, he or she would have about a 75% chance of picking the correct one (Nelson, 1984). Thus, although not perfect, even subjective probability assessments made after one week differentiate correct from incorrect answers at each grain size, and hence should be useful in deciding which answers are least/most in need of "coarsening."

Control process. The results presented so far indicate that participants are able to discriminate between answers that are more likely and less likely to be correct at a given grain size, and suggest that they control the grain size of the chosen answers accordingly. How does the control mechanism operate? Does the nature of the process change over time? Consistent with previous findings in the context of semantic memory reporting (Goldsmith et al., 2002), the participants appear to have relied heavily on subjective confidence in the correctness of their most precise candidate answer when choosing a grain size for their answers. In fact, the within-participant gamma correlations between the assessed-probability-correct of the precise-grain answer on phase 1 of the memory test (henceforth, precise-grain confidence), and the decision to provide that answer rather than the coarse-grain answer (i.e., the choice of grain size) on phase 2 of the test, averaged .91, .88, and .77 for the immediate, 1-day, and 1-week retention intervals, respectively (only the

immediate and 1-week tests significantly differed from each other, $F[1,66] = 8.81$, $MSE = 0.0731$, $p < .01$). This result indicates that there is a very strong relationship between precise-grain confidence and control of grain size at all three retention intervals.

This strong relationship is consistent with a simple satisficing strategy, in which rememberers attempt to provide the most informative (precise) candidate answer whose assessed probability of being correct passes some preset criterion level. To compare this model with alternative models, however, it is necessary to determine whether the control decision might be based instead (or also) on confidence in the coarse-grain answer (henceforth, coarse-grain confidence), or, as implied by a relative-utility model, on the relative disparity between the assessed probabilities of the correctness of the answers at the two candidate grain sizes (i.e., precise-grain confidence/coarse-grain confidence; henceforth, relative confidence). To decide between these models, we conducted a series of logistic regression analyses separately for each retention interval, across all participants and items (see Goldsmith et al., 2002, for a similar approach),[2] with each regression model predicting the likelihood of choosing to provide the precise-grain answer, based on one or more of the confidence measures, just described. These analyses are presented in Table 1.[3]

Model 1 corresponds to the simple correlation between precise-grain confidence and choice of grain size, and again indicates a strong relationship between these two variables. To determine whether there is any further contribution of confidence in the coarse-grain answer, Model 2 includes coarse-grain confidence, and Model 3 includes relative confidence, as additional predictors. The results show that neither coarse-grain confidence nor relative confidence makes any unique contribution to the prediction of the grain choice beyond that of precise-grain confidence (compare the R_L^2 values of Models 2 and 3 with that of Model 1). Moreover, if anything, at least for the immediate and 1-week retention intervals, the regression coefficients for coarse-grain confidence and relative confidence tend to be in the direction opposite to what would be expected under the relative-utility model, with high absolute or relative confidence in the coarse-grain answer increasing the probability that the precise-grain answer, rather than the coarse-grain answer, will be chosen (for a similar result, see Goldsmith et al., 2002). Thus, the results at all three retention intervals are clearly most compatible with the satisficing model, which holds that the choice of grain size should be based primarily on the assessed probability that the precise answer is correct.

Control policy. Our original hypothesis, that participants would maintain a stable level of chosen-grain accuracy over time was based on the additional assumptions (working hypotheses) that both monitoring effectiveness and the participants' control policies would be stable over time. We saw earlier that monitoring effectiveness declined with delayed testing. Was there also a change in the control policy? To answer this question, we estimated the level of the report criterion set by each participant using a procedure adapted from Koriat and Goldsmith (1996b; see also Goldsmith et al. 2002). Essentially, this procedure finds the best fit between the satisficing model and the data of each participant, treating the report criterion value as a free parameter. The report-criterion estimate for each participant is that level of precise-grain confidence which, when applied as the criterion, maximizes the proportion of the participant's grain choices that are correctly predicted by the model (i.e., the proportion of above-criterion answers provided at the precise-grain level and below-criterion answers provided at the coarse-grain level). If a range of criterion values yields an equivalent fit (correct-prediction) rate, the midpoint of the range is chosen.

The mean fit rate yielded by this procedure was quite high overall (89%; comparable to previous results; see Goldsmith et al., 2002), and did not differ for the three retention intervals (90, 88, and 87% for immediate, 1-day, and 1-week tests, respectively), $F < 1$. This again indicates that the satisficing model provides a good description of the participants' grain-control process, and that it is an equally good description regardless of retention interval (cf. earlier logistic-regression results). In addition, the mean criterion estimates were also equivalent for the three retention intervals, averaging .80, .76, and. 83 for the immediate, 1-day, and 1-week

[2] Because the analyses were conducted across participants, it is conceivable that individual differences in overall knowledge or overall confidence might also contribute to the results. However, as in a similar previous analysis (Goldsmith et al., 2002), when such individual differences were partialled out from each analysis (by including each participant's mean accuracy score and mean assessed-probability-correct for the precise-grain and coarse-grain answers in phase 1 as additional predictors), the same pattern of results was obtained. As an additional check, a repeated-measure generalized linear model analysis was conducted using the generalized estimating equation (GEE) method (which corrects for the intra-individual item correlation; Liang & Zeger, 1986; Lipsitz, Laird, & Harrington, 1991). Again, the same pattern of results was obtained.

[3] For the unfamiliar reader, and to clarify the notation, the overall goodness of fit of a logistic regression model is indexed and tested using the G statistic (Hosmer & Lemeshow, 1989), which is analogous to the explained variance (SS_R) in linear regression analyses. An analogue to the proportion of explained variance (SS_R/SS_T) is the R_L^2 statistic (Hosmer & Lemeshow, 1989), which reflects the proportionate reduction in badness of fit relative to the null (intercept-only) model. The interpretation of the standardized coefficients is analogous to their interpretation in linear regression.

Table 1
Results of logistic regression analyses predicting choice of the precise grain size in test-phase 2, on the basis of confidence in the correctness of the answers in test-phase 1, conducted separately for each retention interval

Regression model/analysis		Standardized regression coefficients for independent variables			Model statistics[a]	
No.	Retention interval	Precise-grain confidence	Coarse-grain confidence	Relative confidence[b]	G	R_L^2
1.	Immediate	.76*			352.6	.492
	1-Day	.61*			220.0	.315
	1-Week	.56*			158.5	.270
2.	Immediate	.63*	.16		354.9	.495
	1-Day	.77*	−.19		223.1	.320
	1-Week	.49*	.08		159.0	.271
3.	Immediate	.86*		−.14	354.6	.495
	1-Day	.53*		.10	221.5	.318
	1-Week	.62*		−.14	161.7	.275

Note. All regression models (G statistics) are significant, $p < .001$.
[a] See note 2 for explanation.
[b] Relative confidence = precise-grain confidence/coarse-grain confidence.
* $p < .001$.

tests, respectively, $F(2,69) = 1.55$, $MSE = 0.0220$, $p = .22$. This indicates that the participants' grain-control policy did not change over time. That is, assuming that the participants were following a satisficing model in controlling the grain size of their answers (see preceding section), the control policy did not become more lax or more strict with respect to the minimum assessed-probability-correct required for providing precise answers.

Let us return, then, to the issue of why the accuracy of the chosen-grain answers decreased between the immediate and 1-day retention intervals, before stabilizing between the 1-day and 1-week intervals. Having just seen that the participants were using a similar grain control policy at the different retention intervals, it would appear that the pattern of change in accuracy over time was due primarily to the corresponding pattern of change in monitoring effectiveness. In fact, to achieve a constant level of report accuracy in the face of declining monitoring effectiveness over time, it appears that participants might actually have to raise the report criterion to compensate.

In sum, the results of this experiment indicate that the control of grain size in episodic memory reporting—like the control of grain size in semantic-memory reporting (Goldsmith et al., 2002)—is based on a rather simple satisficing heuristic, and moreover, that the use of this heuristic is stable over time. The consequences for actual memory performance, however, may be rather complex, with constraints on memory accuracy imposed by decreased monitoring effectiveness over time, and perhaps, by restrictions on the coarseness of the answers that can be provided. These conclusions will be examined further, and the restriction on grain size will be loosened, in the next experiment.

Experiment 2

Experiment 2 has two main goals. The first is to examine the control of grain size using a procedure that imposes less restriction on the coarseness of the grain size that is chosen by the participant. Aside from being somewhat more faithful to the type of control that rememberers have over grain size in reporting (quantitative) information in everyday life, this procedure will prevent the inadvertent imposition of artificial constraints on the stability of accuracy over time, and thereby allow a more fair test of participants' ability to achieve high levels of accuracy at delayed testing.

The second goal is to seek more direct support for the claim that the control of grain size in episodic memory reporting is *strategic*. A critic who assumes that hierarchical memory representations have an "all-or-none" character might argue, for example, that rememberers do not strategically choose to provide an answer at a particular grain size. Rather, they simply report the answer at the most precise level of representation that is (still) available in memory at the time of retrieval. By this account, the systematic pattern evidenced in the choice of grain size in Experiment 1 with respect to both accuracy and confidence, might merely reflect the fact that answers that were (still) available at the precise-grain level in the forced-grain phase were more confidently held, more likely to be correct, and more likely to be provided on the subsequent chosen-grain phase, than precise-grain answers that were no longer available in memory on the forced-grain phase, and hence were simply "guessed" in response to experimental instructions. One way to refute such a claim, is to show that the grain size of participants' memory reporting varies in response to changing payoffs for accuracy versus

informativeness. Thus, in the present experiment, we manipulated the relative incentives for accuracy and informativeness, and examined whether the control of grain size is sensitive to this manipulation.

We again tested participants' memory for the quantitative information contained in the fictitious police interview used in Experiment 1. In phase 1 of the test, the participants answered each question at the precise-grain size only, and assessed the probability that their answer was correct. In phase 2, however, the participants were not confined to any specific grain sizes. Instead, they were asked to answer the same questions, using either a precise value (as in phase 1), or specifying an interval of values that they think contains the correct answer. To guide their grain choices, they were told that they would receive a monetary bonus for each correct answer, but that this bonus would be inversely proportional to the grain size of the answer. In addition, the relative payoff for accuracy versus informativeness was manipulated within participants such that for half of the questions, the penalty for a wrong answer was half the amount of the bonus that would be won for a correct precise answer, whereas for the other half the penalty was 10 times as high.

Because the results of Experiment 1 indicated that most of the forgetting of the quantitative information contained in these materials occurs between immediate and 1-day testing, and because significant decreases in chosen-grain accuracy and monitoring effectiveness were found only over that time period, only these two retention intervals were included in this experiment.

Method

Participants

Forty-eight Hebrew-speaking undergraduate students from the University of Haifa participated in the experiment for payment (NIS 35). They were randomly assigned to the two retention intervals.

Materials and procedure

In the first phase, participants in the 1-day test group read the same fictitious interview as in Experiment 1, performed two filler tasks (an attitude questionnaire concerning the Israeli legal system and a figural matrices task consisting of 24 items), and returned one day later for the memory test. Participants in the immediate group read the interview, performed the filler tasks and immediately after were given the memory test.

The memory test, which consisted of the same 22 questions used in Experiment 1, was administered in three phases: a forced, precise-grain phase was followed by a free-grain phase, and finally, by a free-grain confidence phase. In the first of these phases, participants were required to provide a precise value for each question even if they had to guess, and to indicate their confidence in the correctness of each answer using a 0–100% scale that reflects the likelihood that the answer that they provided is correct.

In the second, free-grain phase, participants were given the same set of questions again, but now were given the option to provide either a precise value or a more coarse, bounded-interval answer for each question. They were told that they would win a monetary bonus for each correct answer according to the level of informativeness of that answer. For a correct precise answer they would receive a bonus of NIS 2, but the bonus for a correct interval answer would be less, depending on the width of the interval (the wider the interval, the lower the bonus). No explicit payoff formula was specified for the interval answers. In addition, participants were told that each wrong answer would incur a penalty of either NIS 1 or NIS 10, depending on the question. The applicable penalty was specified next to each question. Two sets of 11 items with approximately equal accuracy rates in Experiment 1 were used for each incentive condition, counterbalanced across participants. To prevent the use of "ridiculously" coarse answers as a means of avoiding the penalty for wrong answers, participants were told that "completely uninformative" answers (for example, that "Yossi and Shay have been friends for 0–100 years") would be considered as incorrect, and incur the applicable penalty for a wrong answer. Participants were assured that if the total penalties exceeded the total bonus, they would not be required to pay their losses; in the worse case they would not receive any bonus for their performance.

After the participants completed the free-grain answering phase, in a third and final phase, they were asked to go over their free-grain answers and assess the probability (0–100%) that each of these was correct (i.e., contains or equals the actual value). They were told that their bonus would not be affected by their confidence ratings.

Results and discussion

As a check on the consistency of the answers between the forced-grain and free-grain phases, inconsistent answers were identified as precise-grain answers that differed between the two phases, and coarse free-grain answers that did not include the forced precise-grain answer as one of the bounded values. The mean percentage of inconsistent answers was only 7%, and this rate did not differ between the two retention intervals. Because the exclusion of inconsistent answers did not change the pattern of any of the reported results, the results are reported with all answers included.

Accuracy and informativeness

Accuracy scores (percent correct) for each participant were calculated for the forced precise-grain answers pro-

vided in test-phase 1, and for the free-grain answers provided under the high and low accuracy incentives in test-phase 2. In addition, a measure of grain size for each answer provided in the free-grain phase was calculated as a logarithmic function of the width of the answer, according to the formula: ln((upper boundary − lower boundary) + 1), a formula which has been shown to roughly capture participants' perception of differences in the informativeness of quantitative-interval answers (Yaniv & Foster, 1995, 1997). By this formula, a precise answer is assigned a grain size of $\ln(1) = 0$, with interval-type answers assigned larger values. The median grain size measure was calculated for the answers provided by each participant in each incentive condition of the free-grain phase. Table 2 presents the mean accuracy and grain-size scores for the participants at the two retention intervals, as well as the mean percentage of precise answers that were provided in the free-grain phase at each retention interval (mean confidence ratings, also appearing in the table, will be addressed later).

First, we note that, as expected, the accuracy of the forced precise-grain answers decreased substantially from immediate testing (64%) to 1-day testing (42%), $F(1,46) = 15.67$, $MSE = 726.55$, $p < .001$. Turning next to the results for the free-grain answers, was the accuracy of these answers stable over the 1-day retention period, despite the decline in precise-grain memory? No it was not. Although overall, the control of grain size did enhance the accuracy of the free-grain answers (67%) compared to the accuracy of the forced precise-grain answers (53%), $F(1,46) = 70.12$, $MSE = 125.52$, $p < .001$, nevertheless, the accuracy of the free-grain answers declined substantially over the 1-day retention interval, from 76% at immediate testing to 58% at the 1-day retention interval, $F(1,46) = 12.64$, $MSE = 627.62$, $p < .001$. In fact, the forgetting rate was no smaller for the free-grain answers than for the forced precise-grain answers, $F(1,46) = 1.24$, $MSE = 125.52$, $p = .27$, for the interaction. Thus, despite the fact that participants were given much greater freedom over the grain size of the reported answers in this experiment than in Experiment 1, and also were given explicit incentives for accurate reporting, free-grain report accuracy was far from stable over the immediate to 1-day retention period.

Was there an effect of accuracy incentive on free-grain accuracy? It can be seen from Table 2 that accuracy incentive did not affect free-grain report accuracy, and there was no interaction between this variable and retention interval (both Fs < 1).

In light of these results, is there any evidence that participants exercised strategic control of grain size in this experiment? Yes, there is: first, as can also be seen in Table 2, a higher proportion of coarse-grain (rather than precise-grain) answers were provided for high-accuracy-incentive items (47%) than for low-accuracy-incentive items (40%), $F(1,46) = 7.96$, $MSE = 156.16$, $p < .01$. A similar difference, though not quite reaching significance, was observed in comparing the proportion of coarse-grain answers provided at the 1-day retention interval (48%) with the proportion provided at immediate testing (38%), $F(1,46) = 2.58$, $MSE = 1044.44$, $p = .06$ (one-tailed). Second, an examination of the grain-size measure, which reflects not only the proportion of coarse versus precise answers, but also the (natural log of) widths of the coarse answers that were provided, shows that significantly coarser answers were provided for high-accuracy-incentive items (0.93) than for low-accuracy-incentive items (0.66), $F(1,46) = 6.07$, $p < .02$, and at 1-day testing (1.10) than at immediate testing (0.49), $F(1,46) = 4.42$, $MSE = 2.0268$, $p < .05$. There was no interaction, $F(1,46) = 1.71$, $MSE = 0.2913$, $p = .20$.

Table 2
Mean report accuracy (percent correct) and confidence (assessed-probability-correct; 0–100%) for forced-grain (phase 1) and free-grain (phase 2) answers in Experiment 2, and mean grain size (Grain; natural log of the interval width) and percentage of coarse grain choices (% Coarse) for answers provided in test-phase 2, as a function of Retention Interval and Accuracy Incentive

Retention interval	Incentive		Forced precise-grain (Phase 1)		Free-grain (Phase 2)			
			Acc.[a]	Conf.[b]	Acc.[a]	Conf.[b]	Grain	% Coarse
Immediate	Low	*M*	64	72	75	82	0.43	34
		SD	24	20	20	19	0.87	22
	High	*M*	64	77	77	83	0.55	42
		SD	21	21	17	21	1.13	21
1-Day	Low	*M*	43	59	58	73	0.89	45
		SD	19	17	20	15	0.87	30
	High	*M*	41	62	57	78	1.31	52
		SD	22	19	25	13	1.12	25

[a] Acc., accuracy.
[b] Conf., confidence.

It appears, then, that participants did exercise strategic control over grain size, increasing (widening) the grain size of their answers in the attempt to increase accuracy: both in response to the relatively large penalty for wrong answers in the high-accuracy-incentive condition compared to the low-accuracy-incentive condition, and—by implication—in response to the decreased accuracy of the precise-grain answers at 1-day testing compared to at immediate testing. However, this control was somewhat limited in terms of its performance consequences: beyond the overall improvement in free-grain accuracy compared to forced precise-grain accuracy, which was achieved equally in both incentive conditions and at both retention intervals, the attempt to increase accuracy further by an additional widening of the answers in response to a particularly strong incentive for accuracy, failed to increase accuracy at all, and simply reduced the overall informativeness (increased the grain size) of the reported answers.

Unlike in Experiment 1, in this experiment there was no inherent constraint on the potential accuracy of the free-grain answers. Therefore, we assume that the source of the limited accuracy improvement resides in the monitoring and control processes underlying the regulation of grain size. We examine each in turn.

Underlying metacognitive mechanisms

Monitoring effectiveness. As in Experiment 1, the effectiveness of the participants' monitoring judgments was evaluated in terms of both calibration and resolution. Calibration curves for the forced-precise answers at immediate and 1-day testing are presented in Fig. 4. With regard to calibration, a comparison of the mean assessed probability correct (confidence) and the actual proportion correct (accuracy) of the forced precise-grain answers (see Table 2, earlier), again indicates that participants were generally overconfident in the correctness of their answers, and that this overconfidence increased over time, with calibration (overconfidence) scores averaging 11% points at immediate testing, and 18% points after a 1-day retention interval, $F(1,46) = 3.97$, $MSE = 187.96$, $p = .05$, for the difference (both means significantly different from zero). Overconfidence also increased over time for the free-grain answers provided in test-phase 2 (not plotted), averaging 6% points at immediate testing, and 17% points after a 1-day retention interval, $F(1,46) = 5.02$, $MSE = 288.71$, $p < .05$, for the difference (both means significantly different from zero). With regard to monitoring resolution, the within-participant gamma correlation between assessed-probability-correct and the actual correctness of each candidate (forced) precise-grain answer averaged .85 on immediate testing and .69 after a 1-day retention interval, $F(1,44) = 3.32$, $MSE = 0.0917$, $p = .08$, for the difference. Thus, as in Experiment 1, considering the results for both calibration and resolution, perhaps part of the decline in free-grain report accuracy on delayed testing may be explained by a decrease in participants' ability to monitor the correctness of potential memory responses.

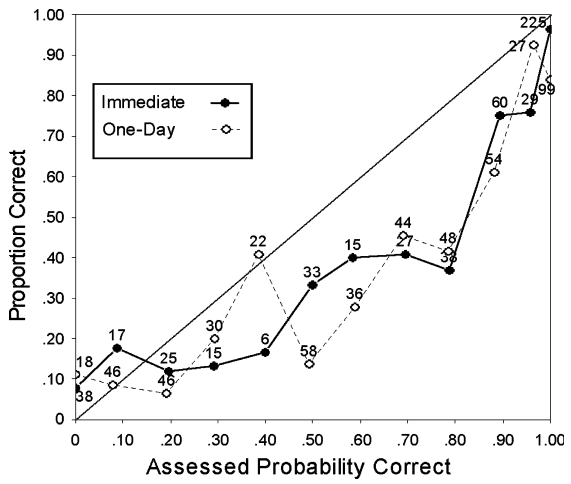

Fig. 4. Calibration curves for the forced precise-grain answers in Experiment 2 (test phase 1), at immediate and 1-day testing. The number of items in each category appears beside each data point.

Control process. Regarding the nature of the control process, as expected, there was again a strong relationship between the assessed probability that a precise answer was correct on the forced-grain phase, and whether or not it, rather than a more coarsely grained answer, was provided on the free-grain phase: the within-participant gamma correlations averaged .85 at immediate testing and .69 at 1-day testing, $F(1,43) = 5.65$, $MSE = 0.0546$, $p < .05$, for the difference (both means significantly different from zero). Similarly, there was also a moderate relationship between the assessed probability that a precise answer was correct on the forced-grain phase, and the (logarithmic) grain size of the answer provided on the free-grain phase: the within-participant gamma correlations averaged $-.56$ at immediate testing and $-.28$ at 1-day testing, $F(1,43) = 8.25$, $MSE = 0.1114$, $p < .01$, for the difference (both means significantly different from zero). Consistent with the satisficing model, these results suggest that the choice of grain size rests heavily on subjective confidence in the correctness of the precise-grain answer, both with respect to whether or not to provide that answer rather than a coarser grained answer, and with respect to the level of coarseness that is chosen—the lower the confidence in the forced precise-grain answer, the larger the grain adjustment of the free-grain answer that is needed to reach (pass) the confidence criterion.

Regarding the control policy, was the same control policy used at each retention interval (as was found in

Experiment 1), and was this policy sensitive—sufficiently so—to the accuracy-incentive manipulation? Using the same procedure as in Experiment 1, report criterion estimates were derived for each participant in each incentive condition (maximizing the proportion of above-criterion answers provided at the precise-grain size and below-criterion answers provided at a coarser-grain size). The results are presented in Table 3. The mean fit rate yielded by this procedure was quite high overall (90.4% correctly predicted grain choices), and did not differ for the two incentive conditions ($F < 1$), though a small difference between the retention intervals approached significance (92.2% vs. 88.6% for immediate vs. 1-day testing, respectively; $F(1,46) = 2.60$, $MSE = 119.44$, $p = .11$). Thus, we again observe a good overall fit between a simple satisficing model and the participants' grain-control decisions. Moreover, the criterion estimates themselves were again equivalent at immediate (.67) and 1-day (.65) testing, $F < 1$, indicating that the grain control policy was stable over time. However, as predicted, the level of the report criterion was sensitive to accuracy incentive, with a higher criterion being adopted for high-incentive items (.73) than for low-incentive items (.60), $F(1,46) = 9.87$, $MSE = 0.0395$, $p < .005$. The interaction between Accuracy Incentive and Retention Interval was not significant, $F < 1$.

This difference in the report criterion estimates for the high- and low-accuracy-incentives yields further evidence that participants exerted strategic control over the grain size of their answers, while also providing some insight into why the effect of this control on actual report accuracy was negligible. Apparently, although the participants did require a higher level of confidence for providing an answer when the penalty for being wrong was relatively high, the extent of the adjustment of the report criterion was too small to have any effect, given the small number of items that were actually affected (an average of only 2.8 items with assessed-probabilities-correct in the range between the low-incentive and high-incentive criterion levels set by each participant), and the limitations imposed by the participants' monitoring effectiveness (see General discussion).

A similar explanation may account for the reduction in memory accuracy over time, despite control over grain size: participants were found to adopt the same report criterion regardless of the retention interval, though, because of decreasing monitoring effectiveness, the adoption of a more conservative report criterion would be needed in order to achieve equivalent accuracy at delayed testing. Overall, then, the results of the present experiment support the claim that participants exert strategic control over the grain size of the answers that they report from episodic memory over time, but indicate that this control is far from optimal.

General discussion

In this article we examined the phenomenon of forgetting over time from a new perspective. Treating the forgetting of episodic details as a starting point, we asked, what can people do, nevertheless, to maintain a reasonable level of report accuracy? One means that rememberers generally have available in real-life memory situations is report option: they can respond "I don't remember" if they feel that they cannot report a piece of information correctly. Another means, which is the focus of this article, is control over grain size: rather than refrain entirely from reporting any information, people may choose to provide an answer at a level of generality at which they are less likely to be wrong.

How might remembers choose the appropriate grain size for their answers in the face of declining memory over time? The theoretical model that guided our investigation assumes that rememberers are (usually) motivated to be both accurate and informative. Therefore, rather than showing indifference to accuracy by always providing very precise answers, or consistently "hedging their bet" by providing very coarse answers, people strategically regulate the grain size of their memory reports—striving to provide as informative (precise) an answer as they can, as long as that answer is judged to be sufficiently likely to be correct. They do so by setting a report criterion for (likely) correctness, and then by monitoring the correctness of candidate answers at different grain sizes, choosing the most precise answer whose assessed probability of being correct passes the report criterion.

Although this "satisficing" model was supported previously in the context of semantic-memory (general-knowledge) reporting (Goldsmith et al., 2002), here we tested the applicability of the model to episodic memory,

Table 3
Mean estimated report-criterion (probability-correct) values and percentages of correct predictions of participants' grain choices for high- and low-accuracy-incentive items at the immediate and 1-day retention intervals in Experiment 2

Retention interval:	Immediate				1-Day			
Accuracy incentive:	Low		High		Low		High	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Criterion value	.60	.25	.74	.25	.60	.34	.71	.28
Percent correct predictions	94	6.3	91	9.3	90	10.5	90	9.4

with a particular focus on the potential contribution of metamemory processes to the stability of memory performance over time. Thus, in addition to examining the change in the amount and quality of the information that is accessible at different retention intervals, we also looked at change and stability in both the grain control policy and in monitoring effectiveness, and how these might jointly account for the pattern of change and stability in overt memory accuracy performance over time.

To guide this examination, we adopted a rather simple set of working hypotheses. Prompted by intriguing findings indicating that under "free-narrative" memory testing, rememberers sometimes maintain high and stable levels of output-bound memory accuracy over time, we hypothesized that by controlling the grain size of their reports, participants might achieve such stability even though both memory for the precise-grain information and memory for coarse-grain information decline over time. They could do so if (a) they are given control over grain size, (b) they adopt a similar report criterion regardless of the retention interval, and (c) their monitoring effectiveness (calibration and resolution) remains stable over time.

How do the results bear on this tentative hypothesis? We first examine the results with respect to the performance consequences of the control of grain size over time, and then turn to the processes underlying the strategic regulation of memory reporting. Following this, we look beyond the horizon of this modest study, and point to some promising directions for future research.

Performance consequences of the control of grain size over time

We examined the performance consequences of the control of grain size across a 1-week retention period in Experiment 1 and across a 1-day period in Experiment 2. Our working hypothesis was that by reporting an increasing proportion of coarse rather than precise answers (Experiments 1 and 2) or by increasing the coarseness (interval widths) of the answers (Experiment 2) at longer retention intervals, participants would be able to achieve relatively stable levels of accuracy over time, despite deteriorating memory representations. The results of the two experiments only partly supported this hypothesis, pointing to some inherent limitations and important variables that must be taken into account.

In Experiment 1, the accuracy of the chosen-grain answers dropped significantly between immediate and 1-day testing, but there was no further decline from 1-day to 1-week testing. In contrast, there was significant forgetting of both precise and coarse forced-grain information across both retention intervals. Thus, the predicted pattern was obtained over the interval from one day to one week, but not between immediate and 1-day testing, where the accuracy decline was diminished—but not prevented—by the control of grain size. As mentioned earlier, it is possible that part of this initial decline was due to the limited control of grain size in that experiment, which constrained the accuracy of the coarse-grain answers. However, given that the drop in Experiment 2 remained significant and substantial, limited grain control is clearly not the whole story. Instead, examination of the underlying monitoring and control processes in both experiments points to the role of these processes: declining monitoring effectiveness in conjunction with a stable control policy appears to be the source of the drop in report accuracy over the initial 1-day retention interval; stable monitoring effectiveness with a continued stable control policy appears to be the source of the stable report accuracy observed over the 1-day to 1-week interval.

These results highlight the critical contribution of metacognitive monitoring and control processes to report accuracy, and raise some interesting questions regarding the role played by control over grain size in the stability or instability of report accuracy over time. Treating the results from this single study with due caution, it appears that under certain conditions control over grain size may be sufficient to enable stable report accuracy over time: across retention intervals in which monitoring effectiveness remains relatively stable, the adoption of a similarly high report criterion at different points in time can yield similarly high levels of report accuracy despite increased forgetting of both precise and coarse information (cf. Fig. 1, earlier). At least for the episodic materials used in this study, monitoring effectiveness appears to achieve such stability within approximately one day after the stimulus event. Whether this time course will generalize to other types of memory situations and materials is, of course, an important question. It is also important to determine whether there might be further reductions in monitoring effectiveness at longer retention intervals. Indeed, it is rather curious that there has hardly been any work addressing how the calibration and resolution of recall monitoring might change over time (but for a study involving recognition memory, see Granhag, 1997).

Can the results help explain the findings of stable accuracy over time in naturalistic recall reporting, mentioned earlier? Perhaps partly. Ebbesen and Rienick (1998), for example, found that although the number of correct statements reported about an experienced event decreased dramatically over a 4-week period, from 1-day to 1- to 4-week testing, the output-bound accuracy of those statements (i.e., the proportion of the statements that were correct; about 90%) remained stable. Essentially the same pattern was obtained by Flin et al. (1992), in comparing the number and accuracy of propositions about a staged event made either one day or five months after the event. Both of these studies

used open-ended questioning procedures that gave participants control over the grain size of their answers,[4] and both involved comparisons that began only after an initial 1-day retention interval had passed. Thus, their findings are broadly compatible with those of the present study. On the other hand, Poole and White (1991) found very high and stable output-bound report accuracy (about 95%) in comparing free narratives about a staged event elicited either immediately after the event or one week later. This stability would appear to have occurred despite a drop in monitoring effectiveness over the initial 1-day period, implied by the current results.

Of course, there are substantial differences between the present study and all of these other studies, not only with respect to the nature of the memory materials, but more importantly perhaps, with respect to the degree of control over memory reporting granted in those (and many other) naturalistic studies. First, the control of grain size in free-narrative and open-ended recall reporting is much more extensive and flexible than was allowed even in the free-grain condition of Experiment 2 here, in which participants were still limited to providing bounded numeric intervals for specific pieces of quantitative information. Second, free-narrative and open-ended recall procedures typically give participants control not only over grain size, but also over report option—allowing them to withhold the pieces of information entirely—which of course, was precluded by the present reporting procedures. Clearly, then, much systematic work remains to be done in clarifying the contribution of personal control over memory reporting to the levels of achieved accuracy over time in naturalistic and real-life remembering.[5]

[4] For example, in explaining their scoring procedure, Ebbesen and Rienick (1998, p. 749), noted that reported information was counted as correct even if it was stated "more generally or specifically" than the actual information. One example that they gave was reporting "car" instead of "Corvette."

[5] It is apparently no coincidence that the results implying high and stable accuracy over time come from naturalistic studies of eyewitness memory and other types of "everyday memory" research. To remain faithful to the conditions of remembering in real-life contexts, such studies often use open-ended questioning procedures that give participants a great deal of freedom in controlling their memory reporting—in deciding what perspective to adopt, how much detail to provide, how much confidence to impart, and even whether to provide an answer at all. Such decisions constitute an important means of regulating the accuracy of memory reporting in real-life contexts (Goldsmith & Koriat, 1999; Koriat & Goldsmith, 1996a). In contrast, traditional laboratory studies have generally denied participants this freedom, instead treating personal control over memory reporting as a methodological nuisance that must be prevented or corrected for (see Goldsmith & Koriat, 1999; Nelson & Narens, 1994).

The strategic regulation of memory reporting

A second issue that we examined concerns the nature of the control process underlying the reporting of episodic information at different grain sizes, and the stability of this process over time. A previous study had addressed the former aspect alone, in the context of semantic memory reporting (Goldsmith et al., 2002). Essentially three alternative models of the grain-control process were examined here in the two experiments.

Experiment 1 compared the "satisficing" model, in which rememberers attempt to provide the most informative (precise) answer whose assessed probability correct passes a preset criterion level, with a "relative utility" model, by which rememberers choose a grain size that maximizes the expected subjective utility of each reported answer. The satisficing model assumes that because of its greater informativeness, the precise-grain answer is treated as the default response; the guiding consideration is whether this response is sufficiently likely to be correct. In contrast, the relative-utility model holds that the assessed probability correct and perceived informativeness of candidate answers at various grain sizes should all play a role. The tendency to volunteer the precise-grain answer should increase to the extent that its assessed probability is relatively close to that of an alternative coarse-grain answer, and to the extent that its perceived informativeness is relatively high compared to that of the coarse-grain answer. The results strongly favored the satisficing model: confidence in the correctness of the precise-grain answer was strongly related to the choice of grain size at all three retention intervals, with increased confidence increasing the likelihood that that answer would be provided. Neither confidence in the coarse-grain answer per se, nor the relative disparity between confidence in the coarse-grain and precise-grain answer, made any additional contribution. Moreover, modeling the operation of the report criterion using confidence in the precise-grain answer alone yielded a very good fit with the data— successfully accounting for about 90% of the participants' grain choices at all three retention intervals.

Experiment 2 compared the satisficing model against a more far-reaching, "non-strategic" alternative, by which the choice of the grain size of a memory response is completely dictated by the state of the memory representation at the time of reporting: the answer will be produced at the most precise representational level that is still available and accessible in memory (if no representation is accessible, the person will need to guess). Although superficially similar to the satisficing model, in this model the rememberer is completely passive: he or she has no freedom in deciding which grain size to provide and there is no involvement of metacognitive monitoring and control processes—confidence in one's answers (assessed probability correct) is merely an epiphenomenon.

The results of Experiment 2 refuted this model by showing that the choice of grain size is influenced by the relative "payoffs" for accuracy and informativeness (which presumably do not change the state of the memory representations). Participants provided more coarsely grained answers when the penalty for being wrong was increased, and this change in reporting behavior was linked to a change in the report criterion: a higher level of confidence was required for reporting precise-grain answers when the penalty for being wrong was high than when it was low. These findings cannot be explained without recourse to a strategic model of grain control, but are explained naturally by the satisficing model, which again yielded a good fit with the participants' choices (about 90% correct predictions)—in both incentive conditions and at both retention intervals.

These results converge with previous findings obtained by manipulating the relative payoff for providing correct fine-grained and coarse-grained answers to general-knowledge questions (Goldsmith et al., 2002, Experiment 3), and with the effects of accuracy incentives (i.e., the penalty paid for wrong answers) on the control of report option (i.e., the withholding of entire answers; e.g., Barnes, Nelson, Dunlosky, Mazzoni, & Narens, 1999; Kelley & Sahakyan, 2003; Koriat & Goldsmith, 1994, 1996b; Koriat et al., 2001). The observed effect of the strategic control adjustments on actual accuracy performance, however, appear to be much more limited for the control of grain size than for report option. Indeed, both in the present study and in Goldsmith et al.'s (2002) earlier study, although the incentive manipulations were clearly influencing the participants' control decisions, the effect on memory performance stemming from these adjustments was negligible.

Why should this be so? The conclusion pointed to by the results of Experiment 2 is that participants' grain adjustments in response to differential incentives are too small relative to the limited resolution of their monitoring judgments. Note that in the context of report option, the moderate-to-high level of monitoring resolution indexed by the within-participant gamma correlations between confidence and the correctness of the individual answers at the precise-grain level, is sufficient for deciding which answers to report and which to withhold—each wrong precise answer that is withheld has a direct positive effect on output-bound accuracy. In contrast, in controlling the grain size of one's answers in order to enhance report accuracy, it is not enough to identify the precise-grain answers that are wrong—one must also be able to determine the (minimum) level of coarseness needed so that the answer will become correct. For example, it is not sufficient to realize that one does not remember the exact number of beers that Benny Sharone drank at the pub; no gain in accuracy will be achieved by widening a wrong precise candidate answer (e.g., "6") to a wrong interval answer (e.g., "between 5 and 7") when the correct answer is "4." Yet this apparently happened quite often in the free-grain phase of Experiment 2: both the proportion of coarse-grain answers and the average grain size of the answers were significantly larger when the penalty for wrong answers was high than when it was low, yet there was no difference between the two incentive conditions in the percentage of answers that were correct (output-bound accuracy).

Why do not participants make larger grain adjustments? One reason is that they are apparently unaware that their grain adjustments are too small, as indicated in both experiments by a substantial degree of overconfidence in the chosen-grain answers, particularly at delayed testing. A converging and startling finding was reported by Yaniv and Foster (1997, Study 2), who had participants provide 95% confidence intervals for estimated quantitative values (e.g., the current population of the US). They found that only 43% of these intervals actually contained the true value. In fact, to reach the targeted proportion correct (95%), the provided intervals would need to be widened on the average by a factor of seventeen! However, Yaniv and Foster (1997) also suggested another reason why people often do not provide wide enough intervals to ensure the accuracy of their answers—they may be reluctant to provide ridiculously coarse answers that violate social-pragmatic norms of communication (e.g., Grice, 1975). Returning to the earlier example regarding the number of beers that Benny Sharone drank at the pub, a participant who feels that she must provide an answer such as "between 1 to 9 beers" in order to be highly confident about the correctness of her answer, may nevertheless be reluctant to do so, because of the implicit prohibition against providing such a ridiculously uninformative answer (in Experiment 2, this prohibition was made explicit in the payoff instructions). We suggest that in most natural memory situations (e.g., a person on the witness stand), a person would prefer to respond "don't know" to such a question (i.e., exercise control of report option), rather than violate communication norms. Perhaps in other situations (e.g., providing answers on an essay exam), however, the possibility of receiving some "points" for a vague answer might dominate the usual tendency to withhold such information (for a discussion of control processes in the context of exam taking, see Budescu & Bar-Hillel, 1993; Koriat & Goldsmith, 1998).

Future directions

The present investigation was predicated on the view that personal control over memory reporting is an intrinsic aspect of everyday remembering (Goldsmith & Koriat, 1999; Koriat & Goldsmith, 1996a). Hence, if we wish to attain a more complete understanding of remembering in real-life contexts, it is important to identify the various types of control and examine their

underlying mechanisms and performance consequences. However, the desire to capture the full richness of real-world memory phenomena is often at odds with the desire to bring the phenomena into the laboratory for controlled experimental investigation (Banaji & Crowder, 1989; Gruneberg & Morris, 1992). In the present study, we tried to achieve an expedient compromise that would offer the benefits of experimental tools and rigor while still tapping some of the fundamental features of the control of grain size in real-world settings. Clearly, though, there are features of the real-life control of grain size that are neglected within our rather artificial experimental paradigm. First, although the operationalization of grain size in terms of bounded intervals for reporting quantitative information is methodologically convenient, it is certainly restrictive. Thus, other forms of control over grain size, will need to be examined, for example, the use of quantitative approximatives and other types of vague linguistic expressions (e.g., "*around* 6 p.m."; Erev, Wallsten, & Neal, 1991; Moxey & Sanford, 1993; Wallsten, 1990; Wierzbicka, 1986), as well as the reporting of information at varying hierarchical levels of abstraction (e.g., reporting "dog" instead of "poodle"; Pansky & Koriat, 2004; Cohen, 2000). Second, the joint control of grain size and report option, typical of open-ended recollection, will need to be investigated. When allowed the option either to provide a coarse-grained answer or to withhold the answer entirely, how do people decide? Will the additional degrees of freedom provided by both types of control of memory reporting allow more effective regulation of memory accuracy and informativeness over time?

More fundamentally, it should be worthwhile to study the metacognitive control of grain size (and report option) in more realistic social contexts, involving a richer set of personal and situational goals. Clearly, in most real-life remembering, the incentives for accuracy and informativeness are not explicit, but rather, are implicit in the personal-social context of remembering (Pasupathi, 2001). Thus, for example, picking up on social cues, people convey more detailed information to attentive than to inattentive listeners (Pasupathi, Stall worth, & Murdoch, 1998), and adjust the detail of their reporting to the perceived needs of the listener (Vandierendonck & Van Damme, 1988; Yaniv & Foster, 1995, 1997). Moreover, the goals of accuracy and informativeness will often be subservient to other goals, such as to be amusing, entertaining, convincing, or impressive (e.g., Neisser, 1988; Sedikides, 1990; Tversky & Marsh, 2000; Wade & Clark, 1993; Winograd, 1994). How do people regulate the grain size of memory reporting in such cases?

The study of the strategic regulation of memory reporting presupposes an expanded conception of memory and memory functioning, in which memory is viewed as a multifaceted tool used in the service of achieving personal and social goals (e.g., Neisser, 1988, 1996; Winograd, 1994). This conception motivates the consideration of a wider range of memory and metamemory processes than does the traditional "storehouse" metaphor of memory (Koriat & Goldsmith, 1996b), and also an examination of these processes within a broader functional context (e.g., Chambres, Izaute, & Marescaux, 2002; Yzerbyt, Lories, & Dardenne, 1998). As Neisser has eloquently argued, remembering is like "doing" (Neisser, 1996), and hence, any complete theory of memory "retrieval" will need to deal with "the reason for retrieval, ... with persons, motives, and social situations." (Neisser, 1988, p. 553).

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic?. *Psychological Bulletin, 93*, 203–231.

Anderson, J. R., Budiu, R., & Reder, L. M. (2001). A theory of sentence memory as part of a general theory of memory. *Journal of Memory and Language, 45*, 337–367.

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist, 44*, 1185–1193.

Barnes, A. E., Nelson, T. O., Dunlosky, J., Mazzoni, G., & Narens, L. (1999). An intergrative system of metamemory components involved in retrieval. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII—Cognitive regulation of performance: Interaction of theory and application* (pp. 287–313). Cambridge, MA: MIT Press.

Begg, I., & Wickelgren, W. A. (1974). Retention functions for syntactic and lexical vs semantic information in sentence recognition memory. *Memory & Cognition, 2*, 353–359.

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review, 10*, 3–47.

Brainerd, C. J., Wright, R., Reyna, V. F., & Payne, D. G. (2002). Dual-retrieval processes in free and associative recall. *Journal of Memory and Language, 46*, 120–152.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology, 2*, 331–350.

Brewer, W. F., & Nakamura, G. V. (1984). The nature and functions of schemas. In R. S. Wyer, Jr. & T. K. Srull (Eds.). *Handbook of social cognition* (Vol. 1, pp. 119–160). Hillsdale, NJ: Erlbaum.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement, 38*, 277–291.

Chambres, P., Izaute, M., & Marescaux, P. J. (Eds.). (2002). *Metacognition: Process, function and use*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Christiaansen, R. E. (1980). Prose memory: Forgetting rates for memory codes. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 611–619.

Cohen, G. (2000). Hierarchical models in cognition: Do they have psychological reality?. *European Journal of Cognitive Psychology, 12*, 1–36.

Conway, M. A., Cohen, G., & Stanhope, N. (1991). On the very long-term retention of knowledge acquired through formal education: Twelve years of cognitive psychology. *Journal of Experimental Psychology: General, 120*, 395–409.

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.

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. (1895/1964). *Memory: A contribution to experimental psychology*. New York: Dover.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*, 519–527.

Erev, I., Wallsten, T. S., & Neal, M. M. (1991). Vagueness, ambiguity, and the cost of mutual understanding. *Psychological Science, 2*, 321–324.

Fisher, R. P. (1996). Implications of output-bound measures for laboratory and field research in memory. *Behavioral and Brain Sciences, 19*, 197.

Flin, R., Boon, J., Knox, A., & Bull, R. (1992). The effect of a five-month delay on children's and adults' eyewitness memory. *British Journal of Psychology, 83*, 323–336.

Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology, 17*, 324–363.

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

Granhag, P. A. (1997). Realism in eyewitness confidence as a function of type of event witnessed and repeated recall. *Journal of Applied Psychology, 82*, 599–613.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.). *Syntax and semantics: Speech acts* (Vol. 3, pp. 41–58). New York: Academic Press.

Gruneberg, M. M., & Morris, P. E. (1992). Applying memory research. In M. M. Gruneberg & P. E. Morris (Eds.), Aspects of memory. The practical aspects (Vol. 1, 2nd ed., pp. 1–17). Florence, KY, US: Taylor and Francis/Routledge.

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

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Hudson, J. A., & Fivush, R. (1991). As time goes by: Sixth graders remember a kindergarten experience. *Applied Cognitive Psychology, 5*, 347–360.

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

Kintsch, W., Kozminsky, E., Streby, W., McKoon, G., & Keenan, J. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior, 14*, 196–214.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language, 29*, 133–159.

Koren, D., Seidman, L. J., Poyurovsky, M., Goldsmith, M., Viksman, P., Zichel, S., & Klein, E. (2004). The neuropsychological basis of insight in first-episode schizophrenia: A pilot metacognitive study. *Schizophrenia Research, 70*, 195–202.

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., & Goldsmith, M. (1998). The role of metacognitive processes in the regulation of memory performance. In G. Mazzoni & T. O. Nelson (Eds.), *Metacognition and cognitive neuropsychology: Monitoring and control processes* (pp. 97–118). Hillsdale, NJ: Erlbaum.

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.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13–22.

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

Lipsitz, S. R., Laird, N. M., & Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika, 78*, 153–160.

Moxey, L. M., & Sanford, A. J. (1993). *Communicating quantities: A psychological perspective*. Hillsdale, NJ: Erlbaum.

Murphy, G. L., & Shapiro, A. M. (1994). Forgetting of verbatim information in discourse. *Memory & Cognition, 22*, 85–94.

Neisser, U. (1984). Interpreting Harry Bahrick's discovery: What confers immunity against forgetting?. *Journal of Experimental Psychology: General, 113*, 32–35.

Neisser, U. (1986). Nested structure in autobiographical memory. In D. C. Rubin (Ed.), *Autobiographical memory* (pp. 71–81). New York: Cambridge University Press.

Neisser, U. (1988). Time present and time past. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.). *Practical aspects of memory: Current research and issues* (Vol. 2, pp. 545–560). Chichester, UK: Wiley.

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

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133.

Nelson, T. O., & & Narens, L. (1994). Why investigate metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: MIT Press.

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

Pasupathi, M. (2001). The social construction of the personal past and its implications for adult development. *Psychological Bulletin, 127*, 651–672.

Pasupathi, M., Stall worth, L. M., & Murdoch, K. (1998). How what we tell becomes what we know: Listener effects on speakers' long-term memory for events. *Discourse Processes, 26*, 1–25.

Payne, B. K., Jacoby, L. L., & Lambert, A. J. (2004). Memory monitoring and the control of stereotype distortion. *Journal of Experimental Social Psychology, 40*, 52–64.

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. (1993). Two years later: Effect of question repetition and retention interval on the eyewitness testimony of children and adults. *Developmental Psychology, 29*, 844–853.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7*, 1–75.

Reyna, V. F., & Kiernan, B. (1994). Development of gist versus verbatim memory in sentence recognition: Effects of lexical familiarity, semantic content, encoding instructions, and retention interval. *Developmental Psychology, 30*, 178–191.

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.

Sedikides, C. (1990). Effects of fortuitously activated constructs versus activated communication goals on person impressions. *Journal of Personality and Social Psychology, 58*, 397–408.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*, 129–138.

Stanhope, N., Cohen, G., & Conway, M. A. (1993). Very long-term retention of a novel. *Applied Cognitive Psychology, 7*, 239–256.

Tversky, B., & Marsh, E. J. (2000). Biased retellings of events yield biased memories. *Cognitive Psychology, 40*(1), 1–38.

Vandierendonck, A., & Van Damme, R. (1988). Schema anticipation in recall: Memory process or report strategy? *Psychological Research, 50*, 116–122.

Wade, E., & Clark, H. H. (1993). Reproduction and demonstration in quotations. *Journal of Memory and Language, 32*, 805–819.

Wallsten, T. S. (1990). The costs and benefits of vague information. In R. M. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 28–43). Chicago, IL: University of Chicago Press.

Wierzbicka, A. (1986). Precision in vagueness: The Semantics of English "approximatives". *Journal of Pragmatics, 10*, 597–614.

Winograd, E. (1994). The authenticity and utility of memories. In U. Neisser & R. Fivush (Eds.), *The remembering self: Construction and accuracy in the self narrative* (pp. 243–251). New York: Cambridge University Press.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science, 2*, 409–415.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition, 25*, 731–739.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy informativeness trade-off. *Journal of Experimental Psychology: General, 124*, 424–432.

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making, 10*, 21–32.

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

Yzerbyt, V. Y., Lories, G., & Dardenne, B. (Eds.). (1998). *Metacognition: Cognitive and social dimensions*. Thousand Oaks, CA: Sage Publications.