THE STRATEGIC REGULATION OF MEMORY ACCURACY AND INFORMATIVENESS

Morris Goldsmith and Asher Koriat

I. Introduction

> When things happen to us, we talk about them. Events do not just happen in words, but that is our primary means of conveying them. When we talk, we do not just recount events one by one in serial order as in a memory experiment.... We tell things differently to different audiences and for different ends. (Tversky & Marsh, 2000, pp. 1–2)

An important development in experimental memory research over the past two decades has been the extension of that research to include phenomena and processes that are characteristic of the richness and complexity of memory in real-life settings. Regardless of the controversies that have accompanied this development, the everyday-naturalistic approach has greatly enriched the study of memory, yielding new experimental paradigms, novel theoretical approaches, and valuable insights.

The central thesis of the present chapter is that particularly in real-life situations, but also to some extent in the laboratory, rememberers strategically regulate the quality and amount of information that they report from memory in accordance with two generally competing goals: accuracy and informativeness. They do so by deciding which items of information to report and which to withhold, and by controlling the level of precision or graininess of the information that they report. These decisions can have a substantial effect on memory performance.

In this chapter, we present a current snapshot of the metacognitive framework that we developed for investigating this regulation, reviewing related work in which some of the essential aspects of the strategic regulation of memory reporting in real-life contexts have been brought into the laboratory for controlled experimental study.

A. Everyday Versus Laboratory Approaches to Memory

Our interest in the strategic regulation of memory reporting stemmed initially from an attempt to clarify some apparent inconsistencies that emerged when comparing laboratory-based findings regarding memory performance with results obtained in naturalistic contexts (Koriat & Goldsmith, 1994). As is well known, there has been a long and sometimes heated debate between proponents of the traditional, laboratory-based study of memory and those who favor the ecological study of memory in naturalistic settings (see, e.g., January 1991 issue of *American Psychologist*). Our analysis of the discussions surrounding this debate (Koriat & Goldsmith, 1996a) revealed three dimensions along which the controversy generally revolved: *what* memory phenomena should be studied (real-life phenomena vs list-learning phenomena), *how* they should be studied (ecological validity vs experimental control), and *where* (real world vs laboratory).

In addition, however, we argued that there seems to be a more fundamental breach underlying these issues that can account for some of the apparent correlation between the "what," "where," and "how" aspects: Underlying the everyday memory approach is a different way of thinking about memory, a different memory metaphor, than that underlying the traditional study of memory. We labeled these metaphors, the *correspondence* and *storehouse* metaphors, respectively. The contrast between the two metaphors provides the metatheoretical foundation for distinguishing two essentially different treatments of memory. As detailed below, in comparison with the traditional, storehouse approach, the correspondence-oriented, everyday approach has engendered (1) an increased focus on the reliability or unreliability of memory in capturing past events, (2) a greater recognition of the active role of the rememberer in controlling memory performance, and (3) a stronger emphasis on the role of subjective-phenomenological experience in remembering.

1. Focus on Accuracy

The traditional laboratory approach to the study of memory has followed Ebbinghaus (1895) in adopting a quantity-oriented conception. In this conception, memory is seen as a storehouse into which discrete items of information are initially deposited and then later retrieved (Roediger, 1980). Memory is then evaluated in terms of the number of items that can be recovered after some retention interval. This approach to memory underlies the traditional list-learning paradigm that continues to produce much of the data that appear in scientific journals.

In contrast, the recent upsurge of interest in everyday memory phenomena implies a different conception of memory. In this conception (following Bartlett, 1932), memory is viewed as a representation or reconstruction of past experience, and hence is evaluated in terms of its faithfulness to past events rather than in terms of the mere number of input items that can be recovered. Embodied in this conception is a *correspondence* rather than a storehouse metaphor of memory (Koriat & Goldsmith, 1996a,b). The correspondence metaphor, with its emphasis on memory accuracy, is apparent in such varied topics as eyewitness testimony, autobiographical memory, spatial memory, memory distortions and fabrications, false memory, memory and metamemory illusions, and schema-based errors. As reviewed in Koriat, Goldsmith, and Pansky (2000), the growing body of work on memory accuracy and distortion has produced a plethora of new paradigms and findings, as well as some specific accuracy-oriented theories that attempt to explain them.

2. *Active Role of the Rememberer*

The interest in everyday memory has led also to a greater emphasis on the functions of memory in real-life contexts and on the active role of the rememberer in putting memory to use in the service of personal goals. Most prominently, Neisser (1996, p. 204) has proposed that remembering should be viewed as a form of purposive action. In his words:

> Remembering is a kind of doing. Like other kinds of doing, it is purposive, personal, and particular: (1) It is *purposive* because it is done with a specific goal in mind; often that goal is to tell the truth about some past event, but on other occasions it may be to entertain, to impress, or to reassure. (2) It is *personal* because it is done by a specific individual and bears the stamp of that individual's characteristic way of doing and telling. (3) It is *particular* because it is done on a specific occasion, in a way that reflects the particular opportunities and demands that the occasion may afford.

Neisser's proposal (see also Winograd, 1994, 1996), together with the idea that memory constructions are "skillfully built from available parts to serve specific purposes" (Neisser, 1996, p. 204), not only promotes a functional perspective in the study of memory but also implies a greater emphasis on self-controlled, regulatory processes in remembering. This emphasis can be seen in an expanded notion of retrieval and remembering (Norman & Schacter, 1996; Winograd, 1996; Koriat, Goldsmith, & Halamish, in press) and in work emphasizing the metacognitive processes of monitoring and control that mediate memory performance (Goldsmith & Koriat, 1999;

Koriat & Goldsmith, 1996b). Complex evaluative and decisional processes used to avoid memory errors or to escape illusions of familiarity have been emphasized by many authors (Burgess & Shallice, 1996; Goldsmith & Koriat, 1999; Kelley & Jacoby, 1996; Schacter, Norman, & Koutstaal, 1998). The operation of these processes is particularly crucial in real-life situations (e.g., eyewitness testimony) in which a premium is generally placed on accurate reporting.

Personal control has not figured prominently in traditional laboratory memory research, perhaps because of its incompatibility with the desire for strict experimental control (Banaji & Crowder, 1989; Nelson & Narens, 1994). Thus, the common approach has been to limit personal control over memory reporting as much as possible (e.g., by using forced-report techniques; Erdelyi & Becker, 1974), or else to attempt to "correct" for it by using techniques such as those provided by the signal-detection methodology (Lockhart & Murdock, 1970) or standard correction-for-guessing formulas (Cronbach, 1984). This approach essentially treats personal control as a methodological nuisance that must be eliminated. However, once we acknowledge that personal control over memory reporting is an intrinsic aspect of real-life remembering (see below), then participants must be allowed such control, but at the same time the underlying dynamics and performance consequences of this control should be systematically investigated.

3. *Emphasis on Subjective Experience*

The focus on memory accuracy and correspondence in real-life remembering has been accompanied by increased interest in the phenomenal qualities of recollective experience. Such qualities have attracted little interest in traditional quantity-oriented memory research. Accuracy-oriented research, in contrast, often involves the assumption that the phenomenal qualities of remembering provide diagnostic clues that are used by rememberers (as well as by observers) for discriminating between genuine and false memories (Conway, Collins, Gathercole, & Anderson, 1996; Koriat, 1995; Ross, 1997). For example, this assumption is central to the source-monitoring framework (Mitchell & Johnson, 2000). In this framework, such properties as perceptual vividness and amount of contextual detail are assumed to help rememberers in specifying the origin of mental experiences. Subjective experience has been examined in connection with autobiographical memories (Brewer, 1992; Conway et al., 1996), false recall (Payne, Jacoby, & Lambert, 2004; Roediger & McDermott, 1995; Schacter, Verfaellie, & Pradere, 1996), post-event misinformation (Zaragoza & Mitchell, 1996), flashbulb memories (Conway, 1995), eyewitness testimony (Fruzzetti, Toland, Teller, & Loftus, 1992), and fluency attributions and misattributions (Kelley & Jacoby, 1998).

In metacognition research, various types of metacognitive feelings, such as the sense of familiarity, the feeling of knowing, and subjective confidence, have been assumed to guide the regulation of search and retrieval processes (Benjamin & Bjork, 1996; Koriat & Levy-Sadot, 1999; Koriat, Ma'ayan, & Nussinson, 2006; Son & Schwartz, 2002). Thus no longer mere epiphenomena, subjective experience is treated as an integral component of the process of remembering (Johnson, 1997; Kelley & Jacoby, 2000; Koriat et al., 2000; Schacter et al., 1998).

B. Competing Goals of Memory Reporting: Accuracy Versus Informativeness

The traditional storehouse metaphor of memory implies a clear goal for the rememberer: to reproduce as much of the originally stored information as possible. This is the essence of the instructions provided to participants in typical list-learning experiments. In contrast, as just discussed, the goals of remembering in everyday life are complex and varied and, in addition, these may be partially or wholly conflicting. Hence, a great deal of skill and sophistication may be required of the rememberer in negotiating between the different goals and in finding an expedient compromise.

In this chapter, we focus on two prominent memory goals that are tied to the storehouse and correspondence metaphors, respectively: quantity, or more generally, informativeness, and accuracy. In real-life situations, these will often be pursued in the service of other, higher-order goals. Importantly, the two goals are generally conflicting. Consider, for example, a courtroom witness who has sworn to "tell the whole truth and nothing but the truth." Even if the witness is sincere in trying to uphold this oath, given the fallibility of memory, it is generally not possible to satisfy both of the implied commitments simultaneously: To avoid false testimony, the witness may choose to refrain from providing information that she feels unsure about. This, however, will tend to reduce the amount of information that she provides the court. Alternatively, she may choose to phrase her answers at a level of generality at which they are unlikely to be wrong. Once again, however, the increased accuracy will come at the expense of informativeness.

In what follows, we present work that examines how rememberers control their memory reporting in the wake of generally competing demands for accuracy and informativeness, and the consequences of this control for their memory performance. Two types of control are considered: The first, *control of report option* (Koriat & Goldsmith, 1994, 1996b), involves the decision to volunteer or to withhold particular items of information. The second, *control of grain size* (Goldsmith, Koriat, & Pansky, 2005;

Goldsmith, Koriat, & Weinberg-Eliezer, 2002), involves choosing the level of precision or coarseness of an answer, when it is provided.

II. The Strategic Control of Memory Reporting: A Metacognitive Framework

In order to bring the essential aspects of the strategic regulation of memory reporting into the laboratory, we adopted an item-based approach that allows the examination of memory quantity and memory accuracy performance within a common framework. In this framework, the two memory properties are distinguished in terms of *input-bound* and *output-bound* measures, respectively (Koriat & Goldsmith, 1994, 1996b). Traditionally, measures of memory performance have been calculated conditional on the input by expressing the number of items recalled or recognized as the proportion or percentage of the total number of items presented. Such measures reflect the amount of presented or studied information that has been retained and is currently accessible. This type of assessment follows naturally from the storehouse metaphor.

Memory performance, however, can also be assessed using *output-bound* measures in which the number of correct items recalled is expressed as a proportion or percentage of the total number of items *reported*. Such measures reflect the *accuracy* of the memory report, in terms of the probability that a reported item is correct. Consider, for example, a participant (witness) who is presented with 25 words (items of information), and in a recall test reports 12 words (provides answers to 12 questions), 10 of which are correct and 2 are commission errors (wrong). Input-bound memory quantity performance in that case is .40 (10/25), that is, 40% of the input-study items have been successfully recalled. In contrast, output-bound memory accuracy is .83 (10/12). That is, 83% of the output-recalled items (answers) are, in fact, correct. This latter measure uniquely reflects the *dependability* of the information that is reported—the degree to which each reported item can be trusted to be correct. Essentially, then, whereas the input-bound quantity measure holds the rememberer responsible for what he or she fails to report, the output-bound accuracy measure holds the person accountable only for what he or she does report.

Importantly, output-bound accuracy and input-bound quantity measures can be distinguished operationally only when rememberers are given the option of *free report*. On forced-report tests, such as forced-choice recognition or (less commonly) forced recall, in which participants are required to provide a substantive response to each and every test item, the input-bound quantity and output-bound accuracy percentages are necessarily equivalent.

This is because the number of output items is the same as the number of input items (Koriat & Goldsmith, 1994, 1996a). For example, if a participant gets 10 out of 25 choices correct on a forced-choice recognition test, we may conclude either that the probability of correctly recognizing an input item is .40 (input-bound quantity) or that the probability that a reported item is correct is .40 (output-bound accuracy). The difference between the two measures is entirely a matter of interpretation—whether one intends to measure quantity or accuracy. In contrast, on free-report tests, such as cued or free recall, participants are allowed to omit items from the memory report or, equivalently, to respond "don't know" if they feel they do not remember an item. In this case, the number of output items may be far fewer than the number of input items.

The option of free report is essential when the focus is on output-bound memory accuracy. Just as an eyewitness cannot be expected to uphold the oath to tell "nothing but the truth" under forced-report conditions, neither does it make sense to hold participants accountable for the errors that they make under such conditions. Indeed, only under free-report conditions, when rememberers have the option to respond "don't know," can we assume that they are actually committed to the accuracy of their memory output. Clearly, in real-life (and most laboratory) settings, rememberers do not simply spew out all of the items of information that come to mind. In fact, as will be seen below, the option to screen out incorrect answers is an important means by which rememberers regulate the quality and quantity of their memory output in real-life settings.

How can the strategic regulation of memory performance in free-report situations be conceptualized and investigated? In searching for a viable research approach, we first turned to signal-detection theory (SDT; Green & Swets, 1966; Swets, Tanner, & Birdsall, 1961). Of course, SDT has been very influential in bringing to the fore the role of subject-controlled processes in memory responding (Lockhart & Murdock, 1970; Norman & Wickelgren, 1969). That framework and its associated Type-1 analyses have been used extensively to investigate the decision processes underlying forced-report recognition memory: Participants in the standard old/new recognition paradigm are assumed to set a response criterion on a continuum of memory strength in order to decide whether to respond "old" (studied) or "new" (foil) to any given test item. Depending on various further assumptions, two indexes are typically derived: a measure of retention, d', and a measure of criterion level, β.

Unfortunately, however, the traditional signal-detection approach (Type-1 analysis) is not very helpful in dealing with the decision process underlying free-report memory performance, that is, with the decision whether to report an answer or to abstain. Therefore, our approach to the problem was to extend the basic logic underlying SDT to free-report situations (as others have done; see Klatzky & Erdelyi, 1985; and see Higham, 2002, for an application of Type-2 SDT analyses, discussed in Section II.D), but also to augment that logic with concepts and methods borrowed from the study of metacognition.

A. The Basic Model: Control of Report Option

Figure 1 presents a simple model of how metamemory processes are used to regulate memory accuracy and quantity performance under free-report conditions (Koriat & Goldsmith, 1996b). The model is deliberately schematic, focusing on the manner in which metacognitive processes at the *reporting* stage affect the ultimate memory performance (cf. the distinction between "ecphory" and "conversion" in Tulving, 1983). Thus, in addition to an unspecified retrieval (or ecphory, reconstruction, and so forth) mechanism,

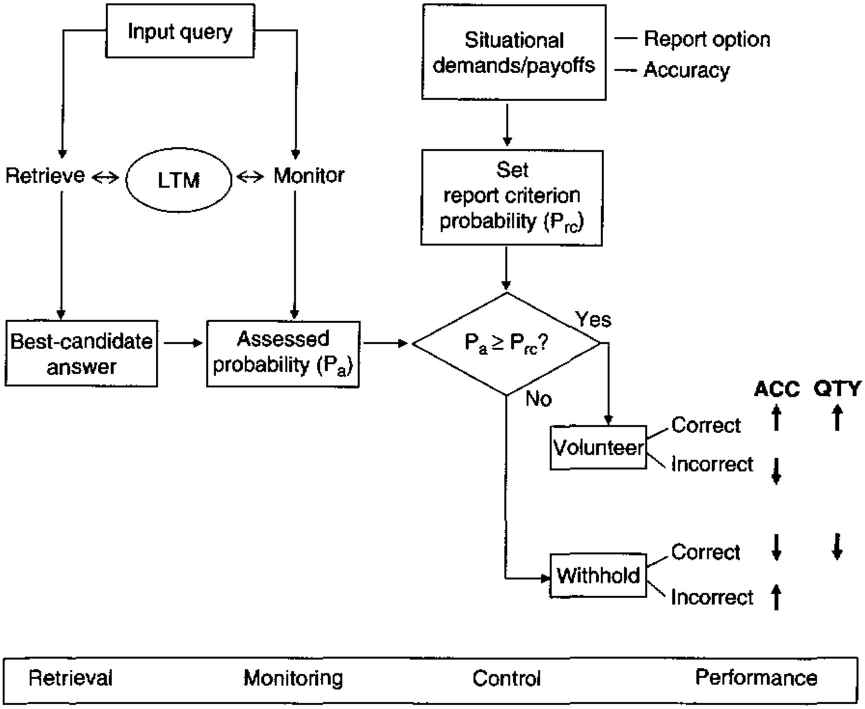

Fig. 1. A schematic model of the strategic regulation of memory accuracy and memory quantity performance, utilizing the option of free report. The upward and downward pointing arrows on the right of the figure signify positive and negative performance outcomes. (Adapted from Koriat & Goldsmith, 1996b.)

we posit a *monitoring* mechanism that is used to subjectively assess the correctness of potential memory responses, and a *control* mechanism that determines whether to volunteer the best available candidate answer (for similar models, see Barnes, Nelson, Dunlosky, Mazzoni, & Narens, 1999; Higham, 2002). The control mechanism operates by setting a report criterion on the monitoring output: The answer is volunteered if its assessed probability of being correct passes the criterion, but is withheld otherwise. The criterion is set on the basis of implicit or explicit payoffs, that is, the perceived gain for providing correct information relative to the cost of providing incorrect information.

Although the model is simple, its implications for memory performance are not. In fact, as will now be explained, within this metacognitive framework, free-report memory performance depends on four contributing factors:

1. *Overall retention*: The amount of correct information (i.e., the number of correct candidate answers) that can be retrieved.
2. *Monitoring effectiveness*: The extent to which the assessed probabilities (subjective confidence judgments) successfully differentiate correct from incorrect candidate answers.
3. *Control sensitivity*: The extent to which the volunteering or withholding of answers is in fact based on the monitoring output.
4. *Report criterion setting*: The report-criterion probability (P_{rc}) above which answers are volunteered, below which they are withheld.

The general assumption is that although people cannot increase the quantity of correct information that they retrieve (e.g., Nilsson, 1987), they can enhance the accuracy of the information that they report by withholding answers that are likely to be incorrect. Hence, the most basic prediction is for a *quantity-accuracy trade-off*: In general, raising the report criterion should result in fewer volunteered answers, a higher *percentage* of which are correct (increased output-bound accuracy), but a lower *number* of which are correct (decreased input-bound quantity). Because raising the report criterion is assumed to increase accuracy at the expense of quantity, the strategic control of memory performance requires the rememberer to weigh the relative payoffs for accuracy and quantity in reaching an appropriate criterion setting.

Of course, this assumes that the participant does in fact volunteer and withhold information on the basis of subjective confidence. This is an assumption that is shared with the SDT framework, but our framework allows for variations in the strength of the relationship between subjective experience and behavior, and treats this as a free parameter in explaining free-report memory performance.

The prediction of a quantity-accuracy trade-off also assumes that the participant's probability assessments are reasonably, but not perfectly, diagnostic of the correctness of the candidate answers. The importance of this assumption has largely gone unnoticed. Indeed, although monitoring effectiveness has attracted much attention among students of metacognition (Koriat, 2007; Metcalfe & Shimamura, 1994; Schwartz, 1994), its performance consequences have only recently begun to be investigated (Barnes et al., 1999; Bjork, 1994; Thiede, Anderson, & Therriault, 2003).

The critical contribution of monitoring effectiveness to both memory accuracy and memory quantity performance emerged in several simulation analyses based on the model (Koriat & Goldsmith, 1996b). Let us assume a testing situation in which 50% of a participant's candidate answers are correct (varying this percentage does not change the basic pattern of results), and manipulate both monitoring effectiveness and report criterion. Figure 2 depicts the accuracy and quantity performance that should ensue under the

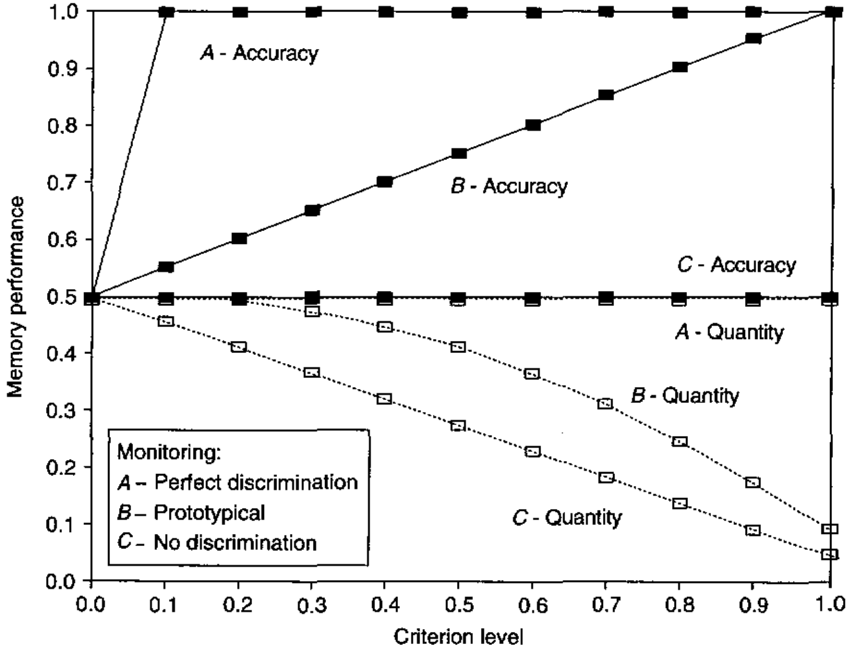

Fig. 2. Simulated memory quantity and memory accuracy performance (proportion correct) plotted as a function of response criterion level, assuming three different levels of monitoring effectiveness (see text for explanation). (Adapted from Koriat & Goldsmith, 1996b.)

model from the use of various report criteria, assuming three different levels of monitoring effectiveness.

Consider first the "prototypical" monitoring condition (Plot *B*). In this condition the participant's confidence judgments are assumed to be uniformly distributed across 11 levels, ranging from 0 (certainly wrong) to 1.0 (certainly right). In addition, these judgments are assumed to be perfectly calibrated, that is, 20% of the answers with confidence (assessed probability) of .20 are correct, 30% of the answers with confidence of .30 are correct, and so forth. Under these conditions, raising the report criterion from 0 (forced report) to 1.0 yields the prototypical quantity-accuracy trade-off: Accuracy increases but quantity decreases as the criterion becomes more strict.

Now, however, consider the plot for the "no discrimination" monitoring condition (Plot *C*) in which the participant's confidence judgments bear no relationship to the actual correctness of the answers. The participant may believe that his or her judgments are diagnostic, but in fact the probability that an answer is correct is .50 regardless of the confidence attached to it. In this extreme case, the participant is unable to enhance his or her memory performance at all by exercising the option of free report: Raising the report criterion does not increase accuracy performance, but simply decreases quantity performance.

Finally, consider the "perfect discrimination" condition (Plot *A*) in which the participant discriminates perfectly between correct and incorrect candidate answers.[1] Here, all correct answers are assigned a subjective probability of 1.0, and all incorrect answers are assigned a probability of 0. In that case, the ideal situation is reached in which the option of free report allows the participant to achieve 100% accuracy with no cost in quantity: For any criterion level greater than 0 (forced report), the participant will volunteer only correct answers and withhold only incorrect answers.

[1] It is important to distinguish between two different indices of monitoring effectiveness, *calibration* and *resolution* (or discrimination accuracy; see, e.g., Lichtenstein et al., 1982; Nelson, 1996; Yaniv, Yates, & Smith, 1991). Calibration captures the absolute correspondence between subjective probabilities and the actual proportions correct. Perfect calibration, however, does not entail perfect monitoring effectiveness at the level of the individual answers. For instance, although a subject may be well calibrated in that, for example, among all items assigned a probability of .60, exactly .60 are correct, this in fact means that the subjective monitoring is not effective enough to differentiate the 60% correct responses from the 40% incorrect responses included in this category. Thus, it is discrimination accuracy (relative correspondence) that is more critical for the effective operation of the control mechanism: When assessed probabilities are polarized between the 0 (certainly wrong) and 1.0 (certainly right) categories, perfect calibration entails perfect discrimination accuracy at the level of individual items. Note, however, that the same discrimination accuracy would be obtained even when the probability values assigned to the two categories were, say, .40 and .41, in which case calibration would be very poor.

These simulations help illustrate the role of two critical factors within the proposed framework: monitoring effectiveness and accuracy motivation. With regard to monitoring effectiveness, clearly some ability to distinguish between correct and incorrect candidate answers is necessary for the control of memory reporting to yield any benefits at all. Moreover, as this ability improves, greater increases in accuracy can be achieved at lower costs in quantity, so that at the extreme, when monitoring effectiveness is perfect, there is no quantity-accuracy trade-off at all.

As far as accuracy motivation is concerned, one can generally increase the accuracy of a memory report by employing a more conservative report criterion. However, under most monitoring conditions, enhancing one's accuracy becomes relatively costly in terms of quantity performance as the criterion level is raised (note the accelerated drop in quantity on the prototypical plot in Fig. 2). Thus, simply giving a person the option of free report may allow a fairly large accuracy improvement to be achieved without much loss of quantity, but placing a larger premium on accuracy should lead to a more serious quantity reduction.

More generally, when considering free-report memory performance, it is both necessary and useful to distinguish between the independent contributions of retention, monitoring, and control. Overall retention (50%, as indexed by forced-report performance at criterion = 0) was the same for all three conditions in Fig. 2. Yet the observed levels of free-report performance could vary dramatically, depending on both the participant's control policy (criterion level) and degree of monitoring effectiveness. We will return to these points again with regard to the empirical results, considered next.

B. Empirical Evidence

Do the monitoring and control processes in fact operate in the postulated manner? To test the basic assumptions of the model, we developed a special two-phase procedure, referred to as the quantity-accuracy profile (QAP) methodology (see Section II.C below). In the first experiment (Koriat & Goldsmith, 1996b, Experiment 1), a general knowledge test was administered to participants in either a recall or a recognition format. The participants initially took the test under forced-report instructions (Phase 1) and provided confidence judgments regarding the correctness of each answer. Immediately afterward, they took the same test again under free-report instructions (Phase 2), with either a moderate accuracy incentive (receiving a monetary bonus for correct answers but paying an equal penalty for wrong answers) or a strong accuracy incentive (in which the penalty was 10 times greater than the bonus).

This procedure enabled us to trace the links postulated by the model (see Fig. 1) between retrieval, monitoring, control, and memory performance:

Retrieval (recall or recognition) was tapped by treating the forced-report answers provided in Phase 1 as representing the participant's best candidate response for each item. Monitoring was tapped by eliciting each confidence judgment as a subjective probability assessment (P_a) associated with each best-candidate answer. This allowed monitoring effectiveness to be evaluated. Control was tapped by examining which answers were volunteered or withheld on Phase 2. This allowed us to determine the sensitivity of the control policy to the monitoring output, and to derive a best-fit estimate of the P_{rc} set by each participant.[2] In addition, a comparison of the estimated report criteria for the two incentive conditions allowed an examination of the predicted effects of accuracy incentive on the participants' control policy. Finally, the design allowed us to evaluate the contribution of monitoring and control processes to the ultimate free-report memory accuracy and memory quantity performance.

The results accorded well with the model. First, participants were found to be fairly effective in monitoring the correctness of their answers. Within-subject Kruskal–Goodman gamma correlations between confidence and correctness (see Nelson, 1984) averaged .87 for recall and .68 for recognition. Second, the tendency to report an answer was strongly tied to confidence in the answer. In fact, the gamma correlations between confidence and volunteering averaged .97 for recall and .93 for recognition! Third, participants who were given the strong accuracy incentive were more selective in their reporting, adopting a stricter criterion than those given the more moderate incentive: They volunteered fewer answers on the average (45%) than did the moderate-incentive participants (52%), and mean confidence for those answers (.93) was higher than those volunteered by the moderate-incentive participants (.84). In addition, the report criterion estimates averaged .84 for the strong-incentive participants versus .61 for the moderate-incentive participants.

Finally, by employing these monitoring and control processes, participants in both incentive conditions were able to enhance their free-report accuracy performance relative to forced report. However, a quantity-accuracy trade-off was observed both in comparing free- and forced-report performance, and in

[2] The procedure for estimating the report criterion (P_{rc}) set by each participant is as follows: For each participant, each assessed-probability-correct (confidence) level from 0 to 1.0 is evaluated as a possible P_{rc}. The model predicts that all items with assessed probability correct greater than or equal to the candidate P_{rc} will be volunteered, and that all other answers will be withheld. The proportion of the participant's actual volunteering and withholding decisions that correspond to the predicted decisions for each candidate P_{rc} are calculated, and the candidate P_{rc} that yields the highest proportion of correctly predicted report decisions (with fit rates generally averaging over 90%) is chosen as the P_{rc} estimate. If a range of values yields an equally good fit, the average of these estimates may be chosen (though in the study referred to here, we used the upper bound instead).

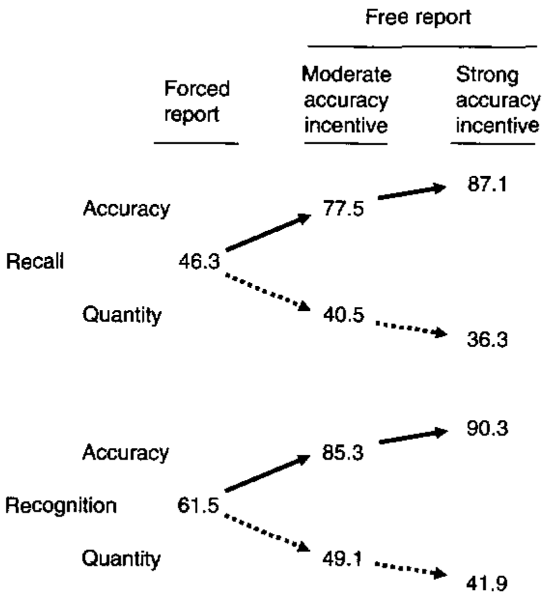

Fig. 3. Results from Koriat and Goldsmith (1996b, Experiment 1). Free-report quantity and accuracy performance (percent correct) as a function of test format (recall vs recognition) and accuracy incentive (strong vs moderate). The means are adjusted for initial differences between the incentive groups in forced-report performance, which is also presented for each test format.

comparing performance under the two incentive conditions (see Fig. 3).[3] Consistent with the simulation analyses, the quantity cost of the improved accuracy increased in relative terms when a higher criterion was employed: Whereas under a moderate accuracy incentive, the option of free report enabled participants to enhance their accuracy substantially at a relatively low cost in quantity performance (a 64% accuracy improvement achieved at a 19% quantity cost for recall; a 33% accuracy improvement achieved at a 26% quantity cost for recognition), the introduction of a stronger accuracy incentive resulted in a further increase in accuracy, but now at a relatively high quantity cost (a further 12% accuracy improvement achieved at a 10% quantity cost for recall; a 6% accuracy improvement achieved at a 15% quantity cost for recognition; based on adjusted means).

A second experiment evaluated the role of monitoring effectiveness (Koriat & Goldsmith, 1996b, Experiment 2). That experiment used the

[3] Due to sampling error, subjects in the high-incentive condition yielded a higher forced-report quantity score (57.3%) than did the moderate-incentive subjects (52.5%). This difference was partialed out in an analysis of covariance to determine the effects of the incentive manipulation of free-report accuracy and quantity performance.

same procedure as in the first experiment (recall and moderate incentive only), but in addition, monitoring effectiveness was manipulated within participant by using two different sets of general knowledge items: One set (the "poor" monitoring condition) consisted of items for which the participants' confidence judgments were expected to be generally uncorrelated with the correctness of their answers (Fischhoff, Slovic, & Lichtenstein, 1977; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Koriat, 1995), whereas the other set (the "good" monitoring condition) consisted of more typical items, for which the participants' monitoring was expected to be more effective. The success of the manipulation can be verified by examining Fig. 4.

Participants based their volunteering decisions heavily on their monitoring output in both monitoring conditions, presumably because they lacked any better predictor. Thus, the gamma correlations between confidence and volunteering averaged .95 and .88 for the good- and poor-monitoring conditions, respectively. More importantly, even when the two sets were matched on retention (by adding some very difficult items to the good-monitoring set) so that *forced*-report performance was equivalent, the good-monitoring condition allowed participants to attain a far superior joint level of *free*-report accuracy and quantity performance: Much better accuracy performance was achieved while maintaining equivalent quantity performance, compared to the poor-monitoring condition (see Fig. 5).

These results, then, reinforce the earlier simulation results in highlighting the criticality of monitoring effectiveness for free-report memory performance. When participants' monitoring effectiveness is good, the option of free report can allow them to achieve high levels of accuracy. In other situations, however, participants' monitoring may be undiagnostic (or perhaps even counterdiagnostic, see Benjamin, Bjork, & Schwartz, 1998) to the point of being useless. Participants still control their memory reporting according to their monitoring output, but the attained level of free-report accuracy may be little better than when participants are denied the option of deciding which answers to volunteer (for similar results using an associative interference manipulation, see Kelley & Sahakyan, 2003; Rhodes & Kelley, 2005).

Of particular importance is the demonstration that monitoring effectiveness can affect memory performance independent of memory "retention" (cf. Fig. 2). Even when retention, as indexed by forced-report quantity performance, was equated across the good- and poor-monitoring subtests in Experiment 2, the joint levels of free-report accuracy and quantity performance were far superior for the good-monitoring subtest than for the poor-monitoring subtest. Clearly, then, free-report memory performance depends on the effective operation of metacognitive processes that are simply not tapped by forced-report performance.

Results from several other studies also suggest a dissociation between monitoring and retention. For example, Kelley and Lindsay (1993) observed

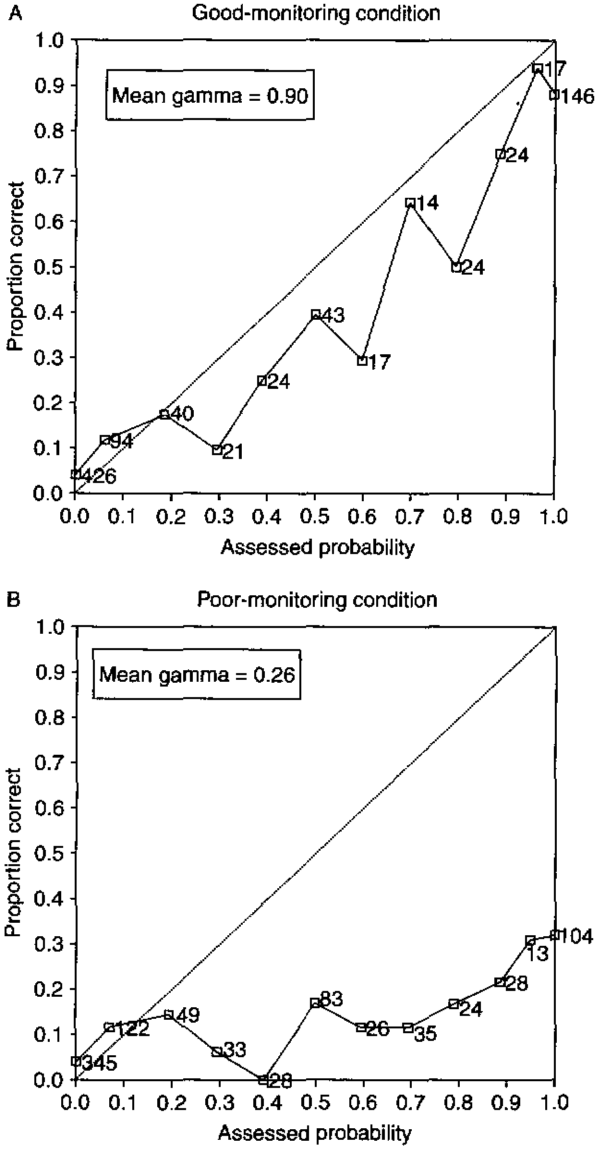

Fig. 4. Calibration plots for the (A) good-monitoring and (B) poor-monitoring conditions in Koriat and Goldsmith (1996b, Experiment 2). The frequency of judgments in each category appears beside each data point, and the mean within-subject gamma correlations between assessed probability correct and actual correctness is also presented.

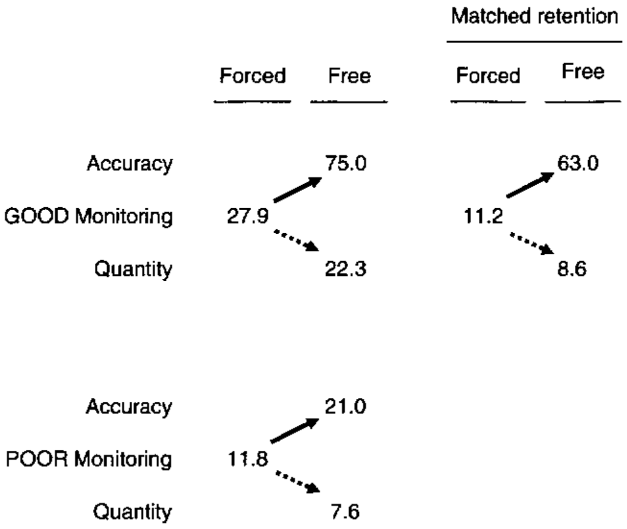

Fig. 5. Results from Koriat and Goldsmith (1996b, Experiment 2). Mean quantity and accuracy performance (percent correct) for the good-monitoring condition, the poor-monitoring condition, and the good-monitoring condition after matching it to the poor-monitoring condition on retention (by including a subset of difficult items).

that advance priming of potential answers to general information questions increased the ease of access to these answers, raising subjective confidence regardless of whether those answers were right or wrong. Similarly, research investigating the cue-familiarity account of the feeling of knowing indicates that feeling-of-knowing judgments can be enhanced by advance priming of the cue, again even when such priming has no effect on actual memory quantity performance (Reder & Ritter, 1992; Schwartz & Metcalfe, 1992). Finally, Chandler (1994) found that exposing participants to an additional set of pictures similar to the studied set increased their confidence ratings on a subsequent forced-choice recognition test, while in fact their actual performance was impaired.

Such dissociations serve to emphasize a basic difference between our proposed framework for conceptualizing the strategic regulation of memory reporting and the well-known (Type-1) SDT approach to memory. Type-1 SDT does not address the separate contributions of memory retention (or memory strength) and monitoring effectiveness to memory performance. In that approach, subjective confidence and memory strength are generally treated as synonymous (Chandler, 1994), and in fact, confidence is often used to *index* memory strength (Lockhart & Murdock, 1970; Parks, 1966; cf. Van Zandt, 2000). Thus, in the forced-report old/new paradigm to which signal-detection methods are typically applied, "control" is isolated in terms of the parameter β, yet "retention" (overall memory strength) and "monitoring effectiveness" (the extent to which the participant's confidence distinguishes "old" from "new" items) cannot be operationally or conceptually separated: Both are equally valid interpretations of d' (Lockhart & Murdock, 1970).

By contrast, in our proposed framework for conceptualizing free-report performance, these latter two aspects (as well as control) are given a separate standing: A person may have effective monitoring, yet very poor retention, or vice versa. Furthermore, poor free-report memory performance, for instance, could derive from poor retention, poor monitoring, an inappropriate control policy, or all three.

The conceptual separation of these components of free-report performance has important implications. At the theoretical level, it calls for more serious efforts to incorporate monitoring and control processes—as well as encoding, storage, and retrieval processes—into our theories and models of memory. At the same time, however, acknowledgment of the potential effects of metamemory processes on memory performance raises an important assessment issue: How should such effects be handled when assessing memory performance? Our approach has been illustrated by the experimental procedure and analyses utilized in the experiments just reported. In the following section, we explicate and expand on this assessment methodology.

C. QAP Methodology

How can one sensibly evaluate a person's memory if memory performance, particularly memory accuracy, is under the person's control? The approach that we developed (Koriat & Goldsmith, 1996b) incorporates metacognitive processes into the assessment of memory performance, while isolating and evaluating their independent contributions to free-report memory quantity and accuracy performance. Thus, rather than deriving a single "point-estimate" index of memory performance, our *QAP* methodology, provides a *profile* of measures that capture various aspects of each participant's cognitive and metacognitive performance in a particular memory task. In addition, the methodology also allows one to examine (by way of simulation) the potential free-report accuracy and quantity performance that participants might achieve, given their particular levels of memory retention and monitoring effectiveness.

The core of the procedure is the two-phase, forced-free paradigm, combined with the elicitation of confidence judgments in the forced-report phase, which was just described in connection with our empirical studies. The role of the forced-report phase is to provide information about memory retention or retrieval which is, as much as possible, unaffected by metacognitive report processes. The role of the free-report phase, beyond that of indicating the actual levels of accuracy and quantity performance that are achieved under free-report conditions, is to provide information about control: the extent to which the report decision is coupled to one's monitoring (control sensitivity), and the strictness or liberality of the report criterion used by the participant (P_{rc}). This is done in conjunction with the confidence judgments that are collected in the forced-report phase. The additional role of the confidence judgments, of course, is to provide information about monitoring per se: its absolute levels, its calibration (e.g., over/underconfidence), and the extent to which it discriminates between correct and incorrect candidate answers (monitoring resolution).

Overall, although the specifics may vary according to one's research goals, our metacognitive free-forced paradigm and associated QAP methodology allow the derivation of up to 10 different measures for each participant (see Table I). In our own work, we have used several variations of the general procedure. Initially, in tasks involving general knowledge questions, we chose to collect the forced- and free-report data in two separate phases, having the participants answer the same set of questions twice: first under forced-report instructions and then again under free-report instructions (or in reverse order). In other experiments, particularly those involving episodic memory tasks, the answers from the initial forced-report phase were carried over to the subsequent free-report phase in which participants simply marked which items they would like to volunteer for points under the specified payoff schedule. Alternatively, however, the free- and forced-report data can be collected on an item-by-item basis, by first forcing the participant to provide an answer, then eliciting a confidence judgment, and finally, having the participant decide whether to volunteer the answer or not (Kelley & Sahakyan, 2003). Each variation of this paradigm has advantages and disadvantages, but the pattern of results obtained across different variations appears to be quite consistent.

An additional component of the methodology and its use within the overall assessment approach still require explication. Similar to the manner in which the plotting of ROC curves in the SDT approach yields additional information beyond what is evident in the parameter values d' and β alone (cf. Higham, 2007), so too with our approach, one can gain additional information by using the answers and associated confidence judgments collected in the forced-report phase to plot the joint levels of free-report quantity and accuracy performance that would ensue from the application of various report criteria to the participants' candidate answers. Like ROC curves, these *QAP curves* (Koriat & Goldsmith, 1996b; see also Goldsmith & Koriat, 1999)[4] allow one to generalize

[4] In our previous presentations of the QAP methodology, it was not entirely clear whether the label "QAP" pertained to our entire methodology or only to the plotted QAP curves. We hope now to correct this problem by adding "curve" or "plot" as a qualifier when referring to this particular component of the overall methodology.

TABLE I

SUMMARY OF MAIN QUANTITY-ACCURACY PROFILE (QAP) MEASURES

Measure	Type	Description	Phase
Retention (or retrieval or ecphory)	Memory	Proportion or percentage of forced-report answers that are correct	Forced
Monitoring resolution (or discrimination accuracy or relative monitoring)	Monitoring	Within-individual gamma correlation between confidence (assessed probability correct) in each answer and the correctness of each answer (Nelson, 1984, 1996), or alternative measures such as ANDI (Yaniv et al., 1991)	Forced
Monitoring calibration (or absolute monitoring) over/underconfidence	Monitoring	Difference between mean assessed probability correct and proportion correct (positive values reflect overconfidence; Lichtenstein et al., 1982).	Forced
Monitoring calibration (or absolute monitoring) squared- or absolute-value deviations	Monitoring	Mean squared- or absolute-value difference between the mean assessed probability correct and proportion correct of each confidence category used in plotting a calibration curve (e.g., Fig. 4; see Lichtenstein et al., 1982)	Forced
Control sensitivity	Control	Within-individual gamma correlation between confidence (assessed probability correct) in each answer and whether or not it was volunteered (see also P_{rc} fit rate)	Forced + Free

Report criterion (P_{rc}) estimate	Control	Estimate of each participant's report criterion (assessed probability level) that yields the maximum fit (fit rate) with his or her actual report decisions (see Footnote 2 for details)	Forced + Free
P_{rc} fit rate (or fit ratio)	Control	The proportion of each participant's actual volunteering decisions that are compatible with the derived P_{rc} estimate, and which is maximized by this estimate (see Footnote 2 for more details). This can also be used as an index of control sensitivity	Forced + Free
Control effectiveness	Control	Absolute value of the difference between the estimated P_{rc} for each participant and the optimal P_{rc}, identified as the P_{rc} level that would maximize the participants' payoff (see Section II.C for details)	Forced + Free
Free-report quantity (input-bound)	Performance	Proportion of correct reported answers out of the total number of questions (or studied items)	Free
Free-report accuracy (output-bound)	Performance	Proportion of correct reported answers out of the number of answers that were volunteered	Free

The table includes typical and alternate names of the measures (in parentheses), the type of cognitive or metacognitive component that they address, a description of how they are calculated, and the source of the experimental data (forced-report or free-report phase) from which they are derived.

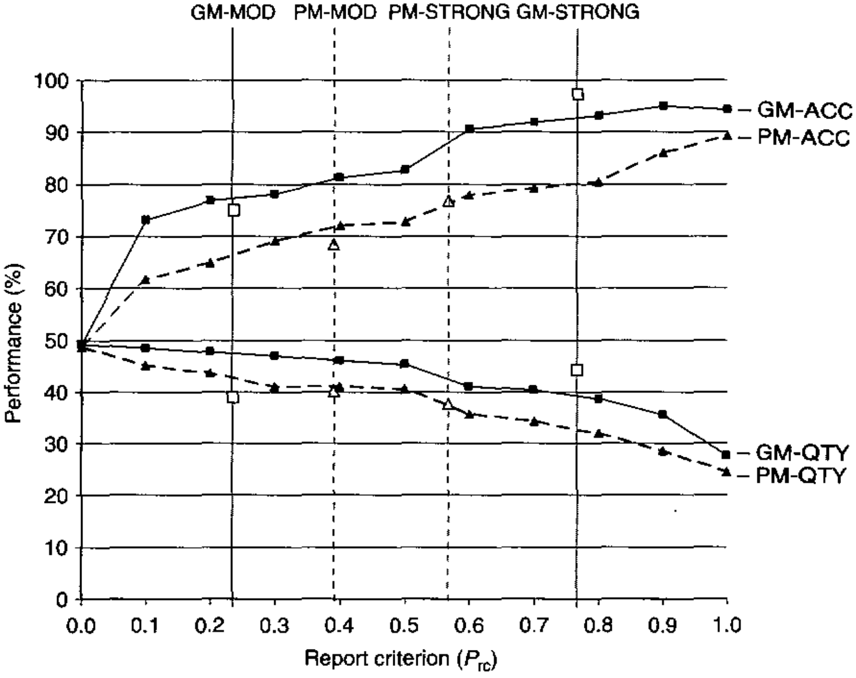

Fig. 6. Illustrative quantity-accuracy profile (QAP) curves for two groups of participants exhibiting different levels of monitoring effectiveness. Potential free-report memory quantity and memory accuracy performance (mean percent correct) is plotted as a function of criterion level for each group. The mean free-report quantity and accuracy scores actually achieved by the participants, subdivided according to the operative level of accuracy incentive (high vs low), are also plotted as open squares or triangles at the point on the *x*-axis corresponding to the criterion estimate for that subgroup. GM, good-monitoring group; PM, poor-monitoring group; ACC, accuracy performance; QTY, quantity performance; STRONG, strong accuracy incentive; MOD, moderate accuracy incentive.

beyond the participants' actual free-report quantity and accuracy performance, to other potential levels of performance, including "optimal" levels (if explicit payoffs for quantity and accuracy have been specified).

To illustrate the method, and how it can be used in conjunction with the other QAP components, in Fig. 6 we present two new QAP curves derived using data from the recall condition of our earlier study (Koriat & Goldsmith, 1996b, Experiment 1). The plots compare the potential quantity and accuracy performance of eight "good-monitoring" participants, those falling in the top quartile of monitoring effectiveness (mean gamma correlation between confidence and correctness of individual items = .95; range: .93–1.0) with eight "poor-monitoring" participants, comprising the bottom quartile of monitoring effectiveness (mean gamma correlation = .75; range: .65–.83).[5] These QAP curves were derived as follows: For each participant, confidence data from the initial forced-report phase were used to calculate the input-bound quantity scores and the output-bound accuracy scores (plotted on the *y*-axis) that would result from the application of 11 different potential report-criterion settings (P_{rc}; plotted on the *x*-axis), ranging from 0 (equivalent to forced report) to 1.0. That is, we assumed that all items with assessed probability correct greater than or equal to each P_{rc} would be volunteered, and calculated the quantity and accuracy scores for that P_{rc} accordingly. The means of these scores at each P_{rc} are plotted separately for each of the two groups of participants. In addition, the actual quantity and accuracy scores achieved by the participants in the free-report phase appear as bullets above the mean estimated criterion level for those participants (based on their actual volunteering behavior; see Footnote 2), subdivided further into those operating under the moderate (1:1 bonus-penalty ratio) versus strong (1:10 bonus-penalty ratio) accuracy incentives, described earlier.

What type of information can be gleaned from these QAP curves? In terms of forced-report performance ($P_{rc} = 0$), the two groups of participants are virtually indistinguishable. Thus, the memory performance ability of the participants in the two groups would be evaluated as "equivalent" under the traditional assessment approach, which often uses forced reporting in attempting to eliminate the contribution of participant-controlled processes. Yet, one can immediately see that as soon as the participants are given the freedom to control their own memory reporting, the higher level of monitoring effectiveness of the good-monitoring participants allows them to achieve substantially better performance than the poor-monitoring participants. In fact, the joint levels of accuracy and quantity performance that can be achieved by the good-monitoring participants are superior to those attainable by the poor-monitoring participants across the range of potential free-report criterion settings ($P_{rc} > 0$). Also, consistent with the results of the earlier simulation analyses, although the poor-monitoring participants can utilize the option of free report to achieve fairly high levels of output-bound accuracy performance, in doing so, they must pay a higher price in quantity performance than do the good-monitoring participants. That is, the good-monitoring participants exhibit a shallower quantity-accuracy trade-off pattern than do their poor-monitoring counterparts. Hence, the memory

[5] Note that the level of monitoring effectiveness exhibited by these "poor-monitoring" participants is still rather good. By comparison, mean gamma was .26 in the poor-monitoring condition of Experiment 2 in that same study (see Fig. 3).

abilities of the participants in the two groups are clearly not equivalent under conditions that allow them to regulate their own memory reporting.

While the QAP curves provide important information about the potential levels of memory accuracy and quantity performance that can be achieved by participants given their specific levels of retention and monitoring effectiveness, they can also be supplemented by information about the actual volunteering policy of the participants under free-report conditions and the actual performance levels that ensue from that policy. This information is derived from the free-report phase of the test procedure. As can be seen, there is a rather good correspondence between the actual free-report quantity and accuracy scores, and the simulated scores based on the forced-report data. The deviations that occur reflect the fact that the participants did not volunteer and withhold their answers entirely in line with the model (the P_{rc} fit rate averaging 93% for this sample) and from the fact that in this illustration, the QAP curves are based on the forced-report data of the participants from both incentive conditions, whereas the actual performance results reflect a subset of those participants, operating under a specific incentive condition.

Examination of the actual free-report performance of the participants brings to the fore the contribution of accuracy motivation: In both monitoring groups, participants strategically regulated their memory reporting according to the operative level of accuracy incentive, with a stricter criterion being adopted when reporting under the strong accuracy incentive than under the moderate accuracy incentive. Interestingly, the good-monitoring participants were much more sensitive to the incentive manipulation than were the poor-monitoring participants, showing a much larger difference in report criteria between the two incentive conditions.

Which group achieved the best actual memory performance, and to what extent can this be attributed to more effective report regulation? Consider the results from the participants in the moderate-incentive condition. Whereas free-report accuracy was higher for the good-monitoring than for the poor-monitoring participants, quantity performance was slightly higher for the poor-monitoring participants. In such a case, one cannot determine, without further assumptions, which level of joint quantity and accuracy performance is better. The answer depends on how much weight is given to accuracy relative to quantity. For example, when testifying on the witness stand in a high-stakes trial, we would tend to give a relatively high weight to the accuracy of the information that is reported. In the early stages of a criminal investigation, however, in order to generate as many leads as possible, one might be more interested in the amount of (correct) information that can be elicited than in the amount of false information (false leads) that is produced. One way of resolving this problem in memory research is to weight the participants' quantity and accuracy performance according to the accuracy incentive payoff schedule under which the participant was operating. By considering the operative payoff schedule, not only do we gain a principled way of combining accuracy and quantity performance into a single performance score, we can also examine the effectiveness of the participants' control of reporting, in particular, the efficiency of the participants in choosing a report criterion that would maximize the "utility" of their memory performance.

Figure 7 presents a set of *payoff curves* corresponding to the QAP plots in Fig. 6, but now describing the potential monetary bonus that could be achieved by the participants in the two monitoring groups under each of the two manipulated payoff schedules (incentive conditions), with the calculations based on the joint levels of quantity and accuracy performance

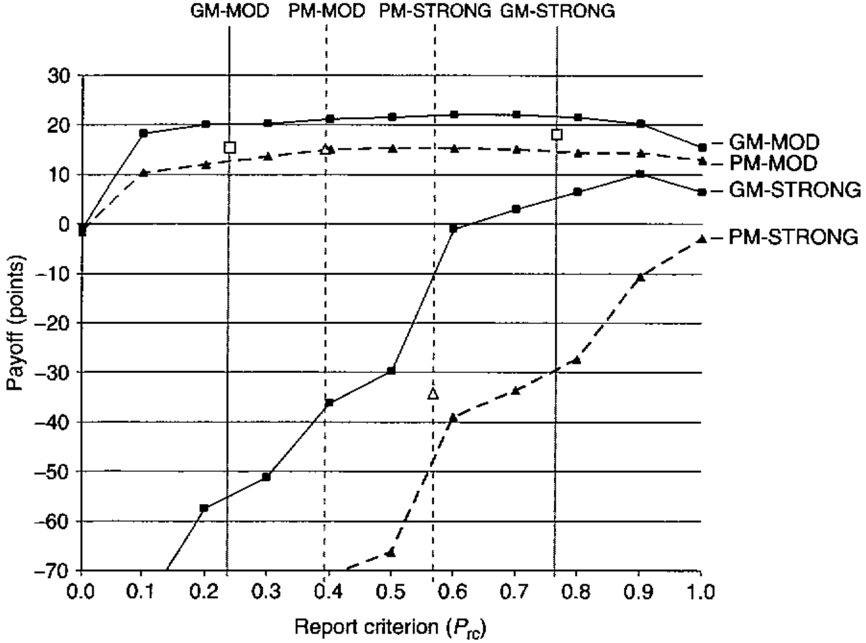

Fig. 7. Illustrative QAP payoff curves corresponding to the QAP performance curves in Fig. 6. Simulated mean free-report point earnings are plotted as a function of criterion level for the participants in each monitoring level × accuracy incentive subgroup. The mean point-payoff actually achieved by the participants, according to the applicable incentive payoff scheme, is also plotted as open squares or triangles at the point on the x-axis corresponding to the criterion estimate for that subgroup. The y-axis is truncated to improve readability. GM, good-monitoring group; PM, poor-monitoring group; ACC, accuracy performance; QTY, quantity performance; STRONG, strong accuracy incentive; MOD, moderate accuracy incentive.

yielded at each potential report criterion level. We see a pattern very similar to the one observed in the preceding analysis: Under each payoff scheme (incentive condition), the potential performance payoff is higher for the good-monitoring group than for the poor-monitoring group across the entire range of possible report criterion settings.

We also see, however, that the actual performance payoff depends on the rememberer's choice of report criterion setting, particularly in the high-incentive condition. In fact, by finding the maximum payoff that could be achieved by each participant and the corresponding criterion setting (based on each participant's individual QAP data), we can identify the *optimal* criterion setting for each participant and the mean optimal criterion for each group or condition. The rememberer's *control effectiveness* can then be evaluated in terms of the difference between the actual payoff and the optimal payoff, and, correspondingly, between the actual (estimated) criterion setting and the optimal criterion setting. Doing this for the specific participants in our illustrative sample, we find that for the good-monitoring participants, the mean optimal criterion settings were .48 and .63 in the moderate- and high-incentive conditions, respectively, compared to the actual (estimated) criterion settings of .23 and .77, respectively. (This cannot be seen in Fig. 7, which combines the data from participants in both incentive conditions.) Thus, these participants' control policy was overly liberal in the moderate-incentive condition and overly conservative in the high-incentive condition, causing them to earn 3 points less than the optimal moderate-incentive payoff and 5 points less than the optimal high-incentive payoff. By comparison, the mean optimal criterion settings for the poor-monitoring participants were .68 and 1.0 in the moderate- and high-incentive conditions, respectively, compared to the actual (estimated) criterion settings of .39 and .57, respectively. Thus, these participants' control policy was overly liberal in both incentive conditions, causing them to earn 6 points less than the optimal moderate-incentive payoff and 29 points less than the optimal high-incentive payoff. Note that the good-monitoring participants were not only more effective in their monitoring, they were also more effective in maximizing their performance by setting an appropriate report criterion than were the poor-monitoring participants, both in terms of the absolute deviation between the actual and optimal criterion settings, and in terms of the difference between actual and optimal payoffs.[6]

The preceding example was designed to illustrate the type of information that can be gained using the QAP assessment approach, and the potential

[6] Note that one can also evaluate the participants' performance with respect to "normative" report criteria by which all (and only) answers with a nonnegative subjective expected value are volunteered. Assuming perfect calibration, these are .50 and .91 in the moderate- and high-incentive conditions, respectively.

utility of evaluating the strategic regulation of memory performance within a decision-theoretic framework. In general, QAP analyses can be used to separate and examine the effects of different variables on memory retention, monitoring, and control in a manner similar to the way Type-1 SDT methods allow one to distinguish differential effects on d' and β. Individual differences and the effects of various factors on the retention and accessibility of information can be examined with respect to forced-report performance. Differences and effects on monitoring effectiveness can be examined in terms of calibration and resolution indexes. Differences in control sensitivity, report criterion, and control effectiveness can also be examined. Finally, the contribution of each of these factors to both actual and potential free-report accuracy and quantity performance can be isolated and compared between conditions or individuals (for further examples of QAP comparisons, see Goldsmith & Koriat, 1999; Koriat & Goldsmith, 1996a,b).

D. QAP or Type-2 SDT?

We have already discussed the similarities and differences between our general approach and the Type-1 SDT approach. However, a variant of our QAP methodology involving Type-2 SDT measures has been put forward by Higham (2002) and used in several subsequent studies (Higham, 2007; Higham & Gerrard, 2005; Higham & Tam, 2005, 2006).[7] It is worthwhile, therefore, to briefly discuss some of the similarities and differences between his methodology and ours (for further discussion, see Higham, 2002).

The psychological model underlying Higham's adaptation of the Type-2 SDT framework to analyze free-report performance is essentially the same as ours. However, some of the performance measures and methods of analysis are different. Like us, Higham assumes the existence of an initial retrieval stage, in which candidate answers are generated, followed by a monitoring and control stage, in which the candidate answers are evaluated and then either volunteered or withheld. As in the QAP methodology, a two-phase,

[7] There has been a great deal of confusion over the years concerning the difference between Type-1 and Type-2 SDT tasks and analyses (for helpful clarifications, see Galvin, Podd, Drga, & Whitmore, 2003; Healy & Jones, 1973). In a Type-1 SDT task, an observer decides which of two events, *defined independently of the observer*, has occurred (e.g., whether a stimulus display contains a target or just noise; whether a presented recognition test item appeared earlier in the study list or not). In a Type-2 SDT task, an observer decides *which of her Type-1 decisions* are correct and which are incorrect. In other words, whereas the Type-1 task taps *cognitive* performance, the Type-2 task taps *metacognitive* performance. Hence, the Type-2 SDT parameters d' (or A') and β (or B''_D) lose their usual Type-1 interpretation: In particular, when the participants' decisions concern their own memory responses, d' (or A') no longer indexes memory, but rather metamemory (monitoring effectiveness). Because Type-2 SDT no longer offers an index of retention cleaned of response bias, an additional (non-SDT) measure must be introduced for this purpose (e.g., forced-report performance; see following discussion).

forced-free procedure is used to gain information about both stages, and as with QAP, the initial retrieval-generation component is indexed in terms of forced-report performance. The monitoring and control components, however, are measured differently. Monitoring effectiveness is measured in terms of the relationship between the free-report decision (volunteer/withhold) and the correctness of the items, calculated using the (Type-2) nonparametric SDT measure, A' (Grier, 1971). Control, or report bias, is measured using the complementary SDT measure, B''_D (Donaldson, 1992), which reflects the tendency to volunteer rather than withhold one's answers, corrected for differences in their overall accuracy.

In the Type-2 approach, then, report control is used to tap monitoring by treating the "volunteer" and "withhold" decisions as reflecting high and low confidence, respectively. In contrast, the QAP methodology taps monitoring directly through confidence judgments. This difference is not arbitrary: Unlike SDT, our framework allows that rememberers may differ in the extent to which they control their memory reporting on the basis of subjective confidence, and that these differences may be interesting sources of variance in free-report memory performance. Therefore, the QAP methodology provides an independent measure of this relationship (control sensitivity). Although the very strong correlations between confidence and reporting obtained with college students, described earlier, might seem to make this dissociation superfluous, in later sections (Sections III.C and III.D) we describe evidence suggesting that control sensitivity may in fact be an important factor to consider in explaining population differences in memory performance.

A second difference concerns the control policy. The B''_D measure used in Type-2 analyses is a measure of report *bias*: For example, if two participants have exactly the same number of correct candidate answers available for reporting (i.e., equivalent forced-report performance), and one of them has a higher volunteering rate than the other, this difference will be reflected in a lower B''_D measure. Note, however, that in terms of our framework, the B''_D measure does not distinguish between the setting of a lower (more liberal) report criterion (P_{rc}), and the alternative possibility, that the increased volunteering rate stems from overconfidence in the correctness of one's answers (i.e., a confidence "distribution shift"; see, e.g., Higham & Tam, 2005, Experiment 2; cf. Wixted & Stretch, 2000). We, in keeping with the decision-making and metacognition literatures (Lichtenstein, Fischhoff, & Phillips, 1982; Nelson, 1996), consider over/underconfidence to be an aspect of *monitoring* rather than of control. This is why in our studies we generally use at least two measures of monitoring effectiveness (one for calibration, and one for resolution; Lichtenstein et al., 1982; Nelson, 1996; see Table I). We reserve the theoretical notions of *control policy* and *report criterion setting* for the idea that independent of how calibrated one is in monitoring the correctness of one's candidate answers, one might be risk-averse, and withhold the answers, or risk-seeking, and volunteer them.

In sum, whereas our theoretical framework and accompanying methodology makes a distinction between five components (and subcomponents) that contribute to free-report performance—retrieval, monitoring resolution (relative monitoring), calibration (absolute monitoring), control sensitivity, and control policy (report criterion or P_{rc})—Higham's Type-2 approach distinguishes only three—retrieval, monitoring resolution, and report bias (over/underconfidence + control policy).

In addition to these measurement differences, however, there also seems to be a fundamental difference between the two approaches in the way in which output-bound memory accuracy is treated. In our approach, output-bound accuracy is on an equal footing with input-bound quantity as a property of interest in its own right. In contrast, Higham's use of the Type-2 SDT approach resembles the traditional use of Type-1 SDT methods, which were generally used to "purify" the measure of memory quantity performance from potential "contaminants" such as response bias. Accuracy (i.e., the false alarm rate) was of interest primarily in order to correct the hit rate for response bias. The same appears to hold for Higham's use of the Type-2 methodology in studying free-report performance. For example, in the studies in which his Type-2 methodology has been applied, output-bound accuracy was not even reported.

Despite these differences, we emphasize that Higham's Type-2 SDT approach has proven to be very valuable in shedding light on the contribution of monitoring and control processes to free-report memory quantity performance (see Section III), and appears to constitute a viable complement to our preferred QAP method, depending on the research context and one's research goals. In the following section, we review work that demonstrates how the general framework that is common to both of our approaches, as well as the more specific assessment methods, can be applied to a variety of research questions and domains.

III. Applications of the Framework

The quantity-oriented research tradition has identified many important variables that strongly affect memory quantity performance. These include study time, massed versus spaced practice, test format (recall vs recognition), depth of encoding, list organization, encoding-retrieval interactions, retention interval, and so forth. The accuracy-oriented approach has focused on other factors such as those involved in producing misinformation effects, reconstructive errors, memory misattributions, and source confusions. Both approaches have also examined individual and population differences. One advantage of our theoretical framework is that it facilitates the merging of issues and findings from the two research traditions, allowing one to examine the effects of various factors on memory performance from both a quantity-oriented and an accuracy-oriented perspective. A second advantage is that armed with the QAP (or Type-2 SDT) methodology, one can not only determine whether a particular factor affects memory accuracy or quantity performance, but also to shed light on the underlying processes that might mediate such effects.

The application of this framework to examine how rememberers use metacognitive monitoring and control processes to regulate the accuracy and quantity of what they report from memory has yielded new insights with regard to several important memory topics and phenomena, such as (1) the effectiveness of different questioning and testing procedures in eliciting accurate memory reports, (2) the credibility of children's witness testimony, (3) memory decline in old age, (4) cognitive and metacognitive impairments related to schizophrenia and psychoactive medication, (5) encoding–retrieval interactions and the encoding specificity principle, and (6) psychometric and scholastic testing. Each of these topics will be considered briefly in turn.

A. The Recall-Recognition Paradox

A prominent variable in memory research is *test format*, recall versus recognition. This variable has been studied in both traditional, quantity-oriented research and in more naturalistic, accuracy-oriented research, with opposing implications: Whereas the general finding from decades of laboratory research (Brown, 1976) is that recognition testing is superior to recall testing in eliciting a greater quantity of correct information from memory, the established wisdom in eyewitness research is that recognition is inferior to recall in eliciting accurate information from rememberers (Hilgard & Loftus, 1979; Neisser, 1988). The latter position stems in part from the belief that directed questioning or recognition testing can have contaminating effects on memory (Brown, Deffenbacher, & Sturgill, 1977; Gorenstein & Ellsworth, 1980; Lipton, 1977). Thus, the general recommendation is to elicit information initially in a free-narrative format before moving on to directed questioning, and, even then, to place greater faith in the former (see Fisher, Geiselman, & Raymond, 1987; Hilgard & Loftus, 1979).

Koriat and Goldsmith (1994), however, showed that this *recall-recognition paradox* actually stems from the common confounding in research practice between test format (recall vs recognition) and report option (free vs forced): Typically, in recognition testing, participants are forced either to choose between several alternatives or to make a yes–no decision regarding each and every item (i.e., forced report), whereas in recall testing participants have the freedom to withhold information that they are unsure about (free report). Comparing performance on a free-recognition test (in which participants had the option to respond "don't know" to individual items) to a free-recall test, Koriat and Goldsmith (1994) found that recognition quantity performance was still superior to recall, but now recognition accuracy was as high or even higher than recall accuracy. Thus, although the superior memory quantity performance of forced-recognition over free-recall testing does appear to stem from the test-format difference (selection superior to production), the generally superior accuracy of free recall over forced recognition appears to stem entirely from report option (free superior to forced).

These initial results were obtained using general knowledge questions, but the same pattern was also observed using a standard list-learning paradigm (Koriat & Goldsmith, 1994, Experiment 2), and in a developmental study, using more naturalistic episodic stimuli (Koriat, Goldsmith, Schneider, & Nakash-Dura, 2001). A subsequent examination of the underlying memory and metamemory components of recall and recognition performance using the QAP procedure (Koriat & Goldsmith, 1996b) indicated that although monitoring effectiveness was in fact somewhat lower for recognition than for recall testing, this disadvantage was more than compensated for by superior memory access and the adoption of a more conservative report criterion under recognition testing.

Based on their results, Koriat and Goldsmith (1994, 1996b) concluded that free recognition may actually be a generally superior testing procedure compared to free recall because it elicits better quantity performance with no reduction in accuracy. This, however, assumes that the option to withhold answers is emphasized by the questioner and clearly understood by the rememberer (Memon & Stevenage, 1996; Pansky, Koriat, & Goldsmith, 2005). In many cases, there may be implicit pressures to respond to directed or recognition queries that cause witnesses to lower their report criterion, even though ostensibly they are given the option to respond "don't know" (see Section III.B).

B. Children's Eyewitness Testimony

The credibility of children's memory, particularly with regard to legal testimony, has been studied intensively in recent years (Bruck & Ceci, 1999; Goodman, 2006). Because of the greater involvement of child witnesses in legal settings, it is important to know whether their recollections of an event can be trusted. Can children be counted on to give a complete and reliable account of past events (to tell the whole truth and nothing but the truth)? This question can be addressed in part in terms of strategic regulatory processes: Are children able to exploit the option of free report to enhance the accuracy of what they report? Can we trust an 8-year-old child, for instance, to effectively censor what she reports, providing only those pieces of information that are likely to be correct? Will her performance be sensitive to specific incentives for accurate reporting? What will be the price in terms of memory quantity? Might differences in the ability and tendency to exert strategic control over memory reporting account for some of the inconsistency in developmental findings, noted earlier?

Results from several studies suggest that children are particularly reluctant to say "don't know" in response to memory questions (Cassel, Roebers, & Bjorklund, 1996; Mulder & Vrij, 1996; Roebers & Fernandez, 2002). Thus, children may be less able or less willing than adults to control their memory reporting on the basis of their subjective monitoring. One approach to correcting this problem is to instruct children in the "rules" of memory reporting. Mulder and Vrij (1996), for example, found that explicitly instructing children aged 4–10 that "I don't know" is an acceptable answer significantly reduced the number of incorrect responses to misleading questions (i.e., questions about events that did not in fact occur). Moston (1987) also found that such instructions induced children aged 6–10 to make more "don't know" responses, but in that study this had no effect on the overall proportion of correct responses. On the other hand, several studies (Cassel et al., 1996; Koriat et al., 2001; Roebers & Fernandez, 2002) have found that children exhibit a greater tendency than adults to provide wrong information from memory even when they are reminded that they have the option to say "don't know."

In our own study (Koriat et al., 2001), 8- to 12-year olds were presented with a narrated slide show depicting an incident on the way to a family picnic, and their memory was later tested under free- or forced-report conditions using a recall or a multiple-choice recognition test format. The results yielded a pattern that was remarkably similar to that described earlier for adult participants. The children's memory accuracy performance was better under free- than forced-report instructions, and the reverse was true for memory quantity performance. For example, in Experiment 1 of that study, memory accuracy increased from 68% under forced-report testing to 81% under free-report testing. In parallel, memory quantity decreased by 5 percentage points. This pattern of quantity-accuracy trade-off, similar to that found with adults (Koriat & Goldsmith, 1994, 1996b), was observed with both younger children (8- and 9-year olds) and older children (11- and 12-year olds). Also, for both age groups, memory accuracy was better under a strong accuracy incentive (88%) than under a moderate accuracy incentive (81%), but here too the improved accuracy was achieved at the expense of quantity performance. Thus, children, even 8 year olds, are capable of exercising the option of free report efficiently to increase the accuracy of their report, and they do so in accordance with the operative level of accuracy incentive. The absolute levels of achieved accuracy, however, differed for the two age groups, with the older children producing more accurate memory reports than the younger children under both the moderate and the strong accuracy incentives, and under both recall and recognition testing.

The main implication of this work is that strategic regulation of memory reporting is a critical factor that can, under the right conditions, allow children to enhance their memory accuracy considerably. Recent work has been directed toward clarifying precisely what those conditions are (Roebers & Schneider, 2005).

C. Memory Impairment in Old Age

Memory decline in old age is both ubiquitous and multifaceted. Here too, the distinction between memory quantity and memory accuracy is crucial. Most research has focused on the decline in the amount of information recalled in old age. Other research, however, indicates an impairment in memory accuracy, with older adults exhibiting greater vulnerability to memory errors and distortions that can have potentially serious consequences such as taking the same medicine twice (Koriat, Ben-zur, & Sheffer, 1988) or being susceptible to scams and con artists (Jacoby & Rhodes, 2006).

There are numerous references in the literature to possible links between the memory impairments associated with aging and those associated with specific neuropsychological deficits, particularly those characteristic of patients suffering frontal lobe lesions (Moscovitch & Winocur, 1995). Neuropsychological evidence suggests that frontal lobes are at least partially involved in metamemory judgments (Janowsky, Shimamura, & Squire, 1989). If old age is associated with impairments in frontal lobe functioning, then we may expect age-related declines in metamemory processes. However, the pertinent results have been mixed and inconclusive (Hertzog & Dunlosky, 2004).

Several studies have focused specifically on old-age-related aspects of metacognitive and neurocognitive functioning that contribute to the strategic regulation of memory performance. Most prominently, Kelley and colleagues (Kelley & Sahakyan, 2003; Rhodes & Kelley, 2005) have utilized our framework and the QAP methodology in conjunction with a clever associative interference paradigm taken from Kato (1985) to compare the strategic regulatory processes of younger and older adults. Kelley and Sahakyan (2003, Experiment 1) found that for control word pairs (not expected to elicit associative interference), although forced-report performance (quantity or accuracy) was superior for younger than for older participants, the older participants utilized the option to withhold answers to narrow the gap between their level of free-report accuracy performance and that of the younger participants. In contrast, for "deceptive" word pairs (in which the retrieval cues evoke a highly accessible associate that competes with the target, thereby presenting a tough challenge to memory monitoring), the age difference in accuracy performance became, if anything, somewhat larger under free report than under forced report. This interactive pattern was accounted for in terms of monitoring effectiveness: First, although both older and younger participants were highly overconfident in the correctness of their responses to the deceptive word pairs, the degree of overconfidence was more pronounced for the older participants. Second, the older participants exhibited lower levels of monitoring resolution for both deceptive and control word pairs.

Additional experiments suggested that the impaired monitoring of the older participants derived from impoverished encoding: When the encoding of the younger participants was disrupted by having them study the word list under divided attention, they exhibited a pattern of performance that was very similar to that of the older participants in terms of both memory accuracy and memory monitoring. Thus, Kelley and Sahakyan suggested that older adults' poorer memory monitoring may derive primarily from their increased reliance on familiarity of candidate responses rather than on recollection of details of the study experience (Jacoby, 1999; Jacoby, Debner, & Hay, 2001), which in turn may derive, at least in part, from poor encoding. A similar conclusion was reached by Rhodes and Kelley (2005), who used the same approach to investigate age differences in memory performance, but now tying these to neuropsychological measures of executive functioning (see also Butler, McDaniel, Dornburg, Price, & Roediger, 2004). In their study, path analyses supported a model in which aging impairs executive functioning, which in turn impairs retention (forced-report performance—a product of both encoding quality and retrieval), which in turn impairs free-report memory accuracy, both directly and by way of impaired monitoring.

Research conducted in our own laboratory (Pansky, Koriat, Goldsmith, & Pearlman, 2002), examining age differences in memory for a short narrated slide show, also indicated an old-age decline in monitoring effectiveness and free-report memory accuracy, and this pattern too was mimicked by a separate group of young adults who watched the slide show under divided attention. In addition, however, we found an interesting age difference in control sensitivity that could not be explained in terms of impaired encoding: Compared to the younger adults in both encoding conditions, the older participants relied less heavily on their confidence judgments in deciding which answers to volunteer and which to withhold under free-report conditions. Moreover, across both age groups, control sensitivity was highly correlated with two measures of executive functioning (with the age factor partialed out): perseverative errors on the Wisconsin Card Sorting Task ($r = -.67$) and the

FAS word fluency test ($r = .46$). These results may perhaps be related to findings implying a breakdown in the relationship between monitoring and control in certain clinical contexts (see Section III.D). This work, together with that of Kelley and colleagues just discussed, points to the need for further investigation of the role of metacognitive monitoring and control processes, and executive functioning in mediating age differences in memory accuracy performance.

D. Clinical Memory Impairment

Metacognitive processes underlying the strategic regulation of performance are also gaining increased attention in research on schizophrenia and on the effects of psychoactive medications. In a series of studies, Koren and colleagues (Koren et al., 2004, 2005; Koren, Seidman, Goldsmith, & Harvey, 2006) adapted our metacognitive framework and QAP methodology to examine the performance of first-episode schizophrenic patients on a metacognitive free-report version of the Wisconsin Card Sorting Task. In that adaptation, patients rated their confidence in each sort and decided whether they wanted that sort to count toward their performance payoff. They found that several of the metacognitive measures from the adapted task correlated more strongly with clinical measures relevant to real-world functioning ("insight into illness" and "competence to consent to treatment") than did traditional neuropsychological measures. Most prominent was control sensitivity, which was more highly correlated with the clinical measures of insight than any of the standard neuropsychological measures that were examined (Koren et al., 2004).

With regard to the strategic regulation of memory reporting, several other results converge in suggesting impaired metacognitive processes in schizophrenia. First, Moritz and colleagues (Moritz & Woodward, 2006; Moritz, Woodward, & Chen, 2006) have observed that even when memory performance is equated, schizophrenic patients exhibit inferior monitoring resolution ("knowledge corruption"), characterized by high confidence in commission errors, compared to healthy controls and to other clinical populations. Second, Danion, Gokalsing, Robert, Massin-Krauss, and Bacon (2001) have used our QAP procedure to compare schizophrenic patients with healthy controls on semantic (general knowledge) memory tasks. In addition to a general deficit in monitoring resolution and calibration (overconfidence), the schizophrenic patients exhibited relatively low control sensitivity (gamma averaging .83 for the clinical patients vs .94 for the healthy controls). Interestingly, similar effects have been found for lorazepam in studies using healthy participants (Massin-Krauss, Bacon, & Danion, 2002). Taken together, these various lines of research indicate two general themes that deserve further attention in future research: (1) a monitoring deficit in which schizophrenic patients are highly confident in wrong responses, and (2) a control deficit, suggestive of an impaired relationship between subjective experience and behavior.

E. Encoding Specificity and Memory Cueing

One of the most basic themes of memory research concerns the critical role of retrieval cues and retrieval-encoding interactions in affecting remembering (Koriat et al., in press). A case in point is the encoding-specificity principle (Tulving & Thomson, 1973), which states that a cue presented during testing will be effective in aiding retrieval to the extent that it has been encoded together with the solicited memory target at study. A large amount of research has provided evidence for this principle (Tulving, 1983) and for the more general idea that retrieval efficiency depends on the extent to which the testing conditions reinstate the overall conditions of study (Schacter, 1990).

Almost all previous research has evaluated the encoding-specificity principle using input-bound memory quantity performance as the criterion (Fisher & Craik, 1977; Morris, Bransford, & Franks, 1977). In contrast, little attention has been paid to the potential effects of encoding-context reinstatement on output-bound memory accuracy performance. Recent work, however, points to the role of metacognitive processes in mediating the effects of context reinstatement on both memory quantity and accuracy performance.

In attempting to clarify the source of context-reinstatement effects, Higham (2002) applied his Type-2 SDT variant of the QAP procedure (see Section II.D) to examine performance in Thomson and Tulving's classic paradigm (Thomson and Tulving, 1970). He replicated the classic finding of superior free-report quantity performance for weak reinstated retrieval cues compared to strong extra-list (unreinstated) cues. In examining the underlying source of this difference, however, he found (somewhat surprisingly) that it was mediated entirely by monitoring effectiveness. Indeed, although retrieval, as indexed by forced-report performance, was equivalent in the two conditions, monitoring effectiveness was much poorer for the strong extra-list cues. For these cues the participants often failed to recognize the targets that they produced, causing them to withhold these items on the free-report phase. A subsequent study, however, in which the cueing strength of the reinstated and extra-list cues was balanced (Higham & Tam, 2006, Experiment 3; see also Zeelenberg, 2005), indicated that the effects of context reinstatement on free-report quantity performance are mediated by both memory retrieval and memory monitoring.

In his studies, Higham did not examine the effects of context reinstatement on memory accuracy. Some insight into these effects and how they are mediated can be gained from an unpublished study by Rosenbluth-Mor (2001). In an adaptation of Tulving and Osler's classic study (Tulving & Osler, 1968), she found that reinstating a weak-associate studied cue at retrieval increased free-report memory quantity performance compared to a baseline (no-retrieval-cue) condition, but had no effect on output-bound memory accuracy. In contrast, providing an extra-list retrieval cue with the same (weak) associative strength to target as the study cue impaired both memory quantity and memory accuracy performance compared to the baseline (no-retrieval-cue) condition. Although preliminary, this pattern suggests that in comparing the reinstated and extra-list cueing conditions, it is not the match between retrieval and study cues that enhances output-bound memory accuracy but rather the mismatch between these cues that impairs accuracy.

With regard to the underlying QAP components, the pattern for forced-report performance mirrored the pattern for free-report quantity performance, suggesting that the cueing effects (both positive and negative) on report quantity were mediated by memory retrieval. At the same time, the pattern for monitoring effectiveness mirrored the pattern for free-report accuracy performance (no monitoring advantage from reinstated cues, but inferior monitoring from extra-list cues), suggesting that the negative effect on report accuracy was mediated by memory monitoring. More specifically, extra-list retrieval cues induced participants to generate a relatively large number of wrong candidate answers with intermediate levels of confidence, compared to the no-retrieval-cue condition, in which the distribution of confidence judgments was more polarized (either one produced the target or else nothing plausible came to mind; cf. Fig. 2).

The preceding work demonstrates how the examination of metacognitive monitoring and control processes can shed new light on the factors affecting both memory quantity and memory accuracy performance in standard, laboratory tasks. Cue reinstatement (encoding specificity) is just one of many classic manipulations and phenomena that might be examined in the context of our metacognitive framework.

F. Psychometric Testing

Many of the standard psychometric tests of intelligence and scholastic aptitude [e.g., the Scholastic Aptitude Test (SAT) and the Graduate Record Examination (GRE) subject tests] use a multiple-choice format in conjunction with *formula-scoring* procedures (Thurstone, 1919) that are designed to discourage guessing and also to correct for it by levying a penalty for incorrect answers, but not for omissions. In fact, the goal of formula scoring is to achieve an estimate of the test-taker's actual knowledge or ability that is "cleansed" from the contribution of guessing (Cronbach, 1984; cf. our earlier discussion of SDT methods). Yet, the penalty for incorrect answers, combined with the option to refrain from answering, effectively puts the test-taker in the position of having to strategically regulate his or her reporting in light of a quantity-accuracy trade-off. Indeed, it is not always clear to test administrators that performance on such tests also taps metacognitive ability, that is, the ability to make effective decisions about whether to risk providing an answer to a question or instead to omit (Budescu & Bar-Hillel, 1993; Koriat & Goldsmith, 1998). Thus, for instance, one test-taker may tend to guess on the basis of even a small amount of partial knowledge, while another may prefer not to provide any answer about which she is unsure (Abu-Sayf, 1979; Gafni, 1990). One test-taker may be effective in distinguishing between answers that are more likely or less likely to be correct, whereas another test-taker may be less effective in discriminating between what she "knows" and what she does not know (Angoff, 1989; Budescu & Bar-Hillel, 1993). Clearly, then, formula scoring is not achieving its intended goal (Albanese, 1988; Angoff & Schrader, 1984; Budescu & Bar-Hillel, 1993; Cross & Frary, 1977; Frary, 1980; Higham, 2007; Slakter, 1968).

Of course, as in the other domains just considered, the fundamental question is not how to get rid of metacognitive contributions to test performance but rather whether we can gain some useful information about the person's abilities from the systematic measurement and analysis of these contributions. Certainly, metacognitive incompetence can have serious consequences. Would we want to certify (or hire the services of) a doctor, lawyer, accountant, psychologist, or engineer who is deficient in discriminating between what she knows and what she does not know (Dunning, Heath, & Suls, 2004), or who, for example, prescribes treatments regardless of whether she is confident of her diagnosis? Would it not be appropriate, then, to include the ability to monitor one's own knowledge and control one's behavior accordingly among those aspects of the examinee's aptitude or achievement that the test is intended to evaluate? Here too, the QAP and Type-2 SDT approaches may allow one to incorporate these components into the psychometric assessment procedure and measure them.

Higham (2007), in fact, has applied his Type-2 SDT approach to the analysis and measurement of the strategic regulation of performance in SAT test taking under formula scoring, with interesting results. In parallel, Notea-Koren (2006) has applied our QAP procedure in the same general context (multiple-choice aptitude test taking) with similar goals and findings. Both studies indicate that the scores of test-takers under formula scoring are affected by the control policy that they adopt and by their level of monitoring effectiveness. In both studies, the test-takers' actual control policies were measured (estimated) and found to differ from an optimal control policy that would maximize their score given their specific level of cognitive performance and monitoring effectiveness (cf. Fig. 7). In addition, results from the Notea-Koren (2006) study show that a component measure of metacognitive ability, monitoring resolution, can contribute unique variance in predicting first-year university grades, beyond the predictive power of the free-report formula score (or the forced-report performance score) alone.

These studies, together with those in the preceding sections, illustrate just a few of the potential domains to which our metacognitive framework for the control of report option, and the QAP assessment methodology, can be extended and applied. In Section IV, we present a further important direction in which the theoretical framework itself has been extended.

IV. Expanding the Framework: Control of Memory Grain Size

The theoretical and empirical work considered so far has focused on how people regulate their memory performance when given the option to withhold individual items of information or entire answers about which they are unsure. Control of report option, however, is just one means by which people can regulate their memory reporting. Indeed, in most real-life memory situations, people do not just have the choice of either volunteering a substantive answer or else responding "I don't know." They also have the option of controlling the "graininess" or level of precision or coarseness of the information that they provide (e.g., describing the assailant's height as "around 6 feet" or "fairly tall" rather than "5 feet 11 inches").

To illustrate, consider a study reported by Neisser (1988), who tested students' memory for events related to a seminar that he taught, using either an open-ended recall format or a forced-choice recognition format. He found the recall format to yield more accurate remembering than the recognition format and noted that this might come as a surprise to memory researchers who are accustomed to the general superiority of recognition testing over recall testing. As discussed earlier (Section III.A), such a finding can perhaps be explained by the effects of report option. Neisser, however, also pointed out a further consideration: Whereas in the recognition format, participants had to make relatively fine discriminations between correct and incorrect response alternatives, in the recall format they seemed to choose "a level of generality at which they were not mistaken" (1988; p. 553).

Along similar lines, Fisher (1996), in assessing participants' freely reported recollections of a filmed robbery, was surprised to find that both quantity performance (number of correct statements) and accuracy performance (output-bound proportion of correct statements) remained constant between two retention intervals across a 40-day span. The anomaly was resolved by considering the grain size of the reported information: Statements made after 40 days contained information that was substantially more coarse (as rated by two independent judges) than the information contained in the earlier statements.

Clearly, then, when rememberers control their own memory reporting, differences in the grain size of the reported information can pose a troubling methodological problem. Here, too, the traditional remedy has been to take control away from the participant, for instance, by using recognition testing or by using stimulus materials, such as word lists, that limit the scope of the problem. Like report option, however, control over grain size is more than just a mere methodological nuisance that needs to be circumvented or corrected for. In most real-life memory situations, it too constitutes an important means by which rememberers regulate the accuracy of their memory reporting and, as such, is an integral aspect of the process of remembering. The challenge is to find a way to systematically investigate this type of control as well. The approach we chose is similar to the one we used for report option and, in fact, assumes a close relationship between these two types of control.

A. Accuracy-Informativeness Trade-Off

Consider a situation in which a witness is asked to answer a set of questions that have to do with quantitative values such as the time of an accident, the speed of a car, the height of an assailant, and so forth.[8] If the witness is forced to answer each question at a specified grain size (to the nearest minute, mile per hour, inch, and so forth), then the accuracy of those answers may be quite poor. However, even though the witness may not remember, say, that the accident occurred precisely at 6:13 pm, she may be able to report that it occurred between 6:00 and 6:30 pm, or perhaps, in the early evening. What, then, will happen if the witness herself is allowed to choose the grain size for her answers? Will she be able to exploit this option in an effective manner, increasing the (output-bound) accuracy of her memory report? On what basis will she choose an appropriate grain size for her answers?

The considerations and mechanisms underlying the choice of grain size in memory reporting appear to be similar to, though somewhat more complex than, those underlying the exercise of report option. Let us

[8] It is methodologically convenient to operationalize grain size in terms of the range or interval width used in reporting quantitative information (Yaniv & Foster, 1995, 1997). We assume that other forms of control over grain size (e.g., vague linguistic qualifiers, "reddish" vs "red") should operate according to similar principles, and in fact recent evidence indicates that they do (Weber & Brewer, in press).

return to the earlier example of a witness who wants to fulfill her vow to "tell the whole truth and nothing but the truth." How should she proceed? On the one hand, a very coarsely grained response (e.g., "between noon and midnight") will always be the wiser choice if accuracy (i.e., the probability of including the true value—telling nothing but the truth) is the sole consideration. However, such a response may not be very informative, falling short of the goal to tell the whole truth. On the other hand, whereas a very fine-grained answer (e.g., 5:23 pm) would be much more informative, it is also much more likely to be wrong. A similar conflict is often faced by students taking open-ended essay exams: Should one attempt to provide a very precise-informative answer, but risk being wrong, or try to "hedge one's bet" by providing a coarser, less informative answer, and risk being penalized for vagueness? In both of these examples, control over grain size can be seen to involve an accuracy-informativeness trade-off similar to the accuracy-quantity trade-off observed with regard to the control of report option.

This idea of an accuracy-informativeness trade-off was brought out nicely by Yaniv and Foster (1995, 1997) in the context of judgment and decision making. They showed that when people are asked to give quantitative estimates for the purpose of decision making, they tend to consider the recipient's desire to obtain a useful response (cf. Grice, 1975), and often sacrifice accuracy for informativeness. Recipients of information generally require estimates that are both sufficiently informative for their current needs and appropriately accurate. For example, information that the inflation rate will be "between 0% and 80%" in the coming year will not be appreciated by the recipient, although it is likely to be correct. In fact, Yaniv and Foster (1995) found that to some extent, recipients of information actually prefer a somewhat inaccurate but precise-informative estimate (e.g., that the inflation rate will be "5–6%" when it turns out to be 7%) to an overly coarse, uninformative estimate that is "technically" correct.

B. Satisficing Versus Utility-Maximizing Models of Grain Control

How does one find an appropriate compromise between accuracy and informativeness in choosing a grain size for his or her answers? One simple strategy that we considered is to provide the most finely grained (precise) answer that passes some preset report criterion (in terms of assessed probability correct). Thus, for example, our earlier witness might try to answer the question to the nearest minute, to the nearest 5 minutes, 10 minutes, 15 minutes, and so forth, until she is, say, at least 90% sure that the specified answer is correct. Goldsmith et al. (2002) called this the *satisficing model* (cf. Simon, 1956) of the control of grain size: The rememberer strives to provide as much information as possible, as long as its assessed probability of being correct satisfies some reasonable minimum level. Note that this model is similar to the one presented earlier with regard to report option: As with report option, the assessed probability correct of each answer that is volunteered must pass a report criterion, and the setting of the criterion level should depend on the relative incentives for accuracy and informativeness in each particular situation.

A more complex, alternative model was also examined. According to the *relative expected-utility maximizing model*, rememberers monitor in parallel the likely correctness (assessed probability correct) of candidate answers at various grain sizes, and evaluate the informativeness (subjective value or utility) of the answer at each grain size. Combining the outputs of these two operations, they then calculate the subjective expected value or utility of the answer at each grain size (e.g., assessed probability correct × subjective value or utility), compare these values, and choose the answer that maximizes the subjective expected value or utility. Such a relative comparison process, while aiming for a more optimal grain-choice solution than the satisficing model, would seemingly place a much heavier cognitive and metacognitive burden on the rememberer than does the satisficing model.

Before turning to the empirical evidence with respect to these models for the control of grain size, note that although they differ in their specifics, they share a common conception of the choice of grain size as being based on two metacognitive processes: (a) a monitoring process that assesses the probability that answers at different grain sizes are correct, and (b) a control process that uses the monitoring output, together with other information (e.g., the perceived informativeness of the answers, and/or the relative incentives for accuracy and informativeness) in order to decide on the appropriate grain size for a particular answer.

C. Empirical Evidence

Goldsmith et al. (2002) conducted a systematic study of the control of grain size in reporting from semantic memory. The main goal of that study was to determine whether the general metacognitive framework of monitoring and control that had been developed earlier to address the control of report option, would be useful in studying the control of grain size as well.

In that study, participants answered a set of general knowledge questions, all of which related to quantitative-numeric information: time, date, age, distance, speed, and so forth. The questions were presented in two phases: In the first phase, participants gave their best answer to each item using two different bounded intervals (grain sizes), the widths of which were specified by the experimenter. For example, "When did Boris Becker last win the Wimbledon men's tennis finals? (A) Provide a 3-year interval; (B) Provide a 10-year interval." The two grain sizes were tailored for each item such that the coarse-grained answer specified a relatively wide interval, that would yield a mean proportion correct of about 75%, whereas the fine-grained answer specified a more narrow interval (or in some cases a specific value, e.g., year), that would yield a mean proportion correct of about 30%. In the critical second phase, the participants went over their answers, and for each item, indicated which of the two answers (i.e., which of the two grain sizes) they would prefer to provide, assuming that they were "an expert witness testifying before a government committee."

In Experiment 1 of the study, participants chose to provide the fine-grained answer in about 40% of the cases, implying that the choice of grain level was not guided solely by the desire to be correct (in which case they would have always chosen the coarse-grained answer), nor solely by the desire to be informative (in which case they would have always chosen the more precise, fine-grained answer). Instead, the participants tended to choose the coarse-grained answer when the more precise answer was deemed too unreliable: Answers that the participants chose to provide at the fine-grained level had a relatively high (about 50%) chance of being correct, whereas the fine-grained answers that they would have provided, had they not chosen to provide the coarse-grained answer instead, had a relatively low (about 20%) chance of being correct. Moreover, by sacrificing informativeness in this strategic manner, the participants improved their overall accuracy substantially (to about 60%) compared to what they would have achieved by providing the fine-grained answers throughout (about 30%). The maximum accuracy that could have been achieved by providing only coarse-grained answers was somewhat higher (about 75%).

In Experiment 2 of the study, the collection of confidence judgments (assessed probability correct) for the answers at each grain size in the first phase of the design helped shed light on the monitoring and control processes underlying the choice of grain size. First, with regard to monitoring, the participants were fairly successful in discriminating between correct and incorrect answers at each grain size, with moderately high within-participant gamma correlations between confidence and correctness of each answer (averaging about .50) for both the fine-grained and the coarse-grained answers. Second, with regard to control, there was a strong relationship between confidence in the fine-grained answer and choice of grain size, with within-participant gamma correlations between confidence in the fine-grained answer and the choice to provide that answer (rather than the coarse-grained answer) averaging about .80.

In order to gain more insight into the process underlying the regulation of grain size, we conducted some further analyses. According to the satisficing model, participants should provide the fine-grained answer if its assessed probability passes the report criterion, otherwise they should give the coarse-grained answer. Therefore, confidence in the fine-grained answer should be the primary predictor of grain choice, whereas confidence in the coarse-grained answer should be irrelevant. In contrast, according to the expected-utility maximizing model, both of these should contribute to the grain choice. In particular, because the expected subjective utility of providing the coarse-grained answer increases as confidence (assessed probability correct) in that answer increases, all else equal, there should be a positive relationship between confidence in the coarse-grained answer and the tendency to provide that answer (rather than provide the fine-grained answer). The results of several multiple (logistic) regression analyses clearly favored the satisficing model: When both confidence in the fine-grained answer and confidence in the coarse-grained answer were included as joint predictors of the choice of grain size, confidence in the fine-grained answer was the primary predictor, with a standardized regression coefficient over three times as large as the coefficient for coarse-grained answer confidence. Moreover, the sign of the coefficient for coarse-answer confidence was in the opposite direction from what would be predicted by the expected-utility maximizing model: Holding confidence in the fine-grained answer constant, confidence in the coarse-grained answer showed a weak but significant *negative* relationship to choice of the coarse-grained answer.

Finally, in a third experiment, the setting of the report criterion was shown to be strategic. In that experiment, the relative weight assigned to informativeness versus accuracy was manipulated by introducing explicit monetary incentives for correct answers at the two grain sizes in the second phase: A higher bonus was paid for correct fine-grained answers than for correct coarse-grained answers, and this ratio was 2:1 for half of the items (weak informativeness incentive) and 5:1 for the other half (strong informativeness incentive), counterbalanced across participants. As predicted, more fine-grained answers were provided in the strong informativeness-incentive condition than in the weak incentive condition, decreasing the accuracy of those answers, as well as the average confidence in those answers. The results involving confidence were again consistent with the satisficing model and inconsistent with the expected-utility maximizing model. Using a procedure similar to the one described earlier as part of the QAP methodology, the report criterion set by each participant for providing fine-grained answers was estimated, with the mean estimated criterion significantly more liberal in the strong informativeness-incentive condition (.58) than in the weak-incentive condition (.74).

D. Control of Grain Size in Episodic Memory Reporting Over Time

As in the case of report option, a consideration of the control of grain size in memory reporting has begun to shed light on other memory phenomena and issues. One example is the potential role of control over grain size in modulating the changes that occur in memory over time. Goldsmith et al. (2005) examined the regulation of report grain size over different retention intervals. Starting from the well-known finding that people often remember the gist of an event though they have forgotten its details (e.g., Kintsch, Welsch, Schmalhofer, & Zimny, 1990; Koriat, Levy-Sadot, Edry, & de Marcas, 2003), they asked whether rememberers might exploit the differential forgetting rates of coarse and precise information in regulating the accuracy of the information that they report over time.

Consider Neisser's and Fisher's anecdotal observations regarding the control of grain size, mentioned earlier (Fisher, 1996; Neisser, 1988). The general hypothesis implied by these observations is that in recalling episodic information from memory, rememberers 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 findings indicating that detailed information suffers a faster forgetting rate than coarse information (Kintsch et al., 1990; Koriat et al., 2003; Reyna & Kiernan, 1994), 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, Budiu, & Reder, 2001; Brainerd, Wright, Reyna, & Payne, 2002; Koutstaal, 2006).

In order to test this hypothesis, Goldsmith et al. (2005, Experiment 1) had participants read a short transcript describing fictitious events surrounding a bar-room argument, and tested their memory for these events either immediately, or after a one-day or one-week retention interval. The experimental paradigm was similar to that used in Goldsmith et al. (2002) except for the episodic nature of the memory material. Again, the test questions pertained to various items of quantitative information (heights, weights, ages of the characters, times of day, distances, etc.), and were initially answered at two fixed grain sizes, one precise (a specific value) and the other coarse (a bounded interval). In the second, grain-choice phase, the participants were instructed to choose for each item, the answer (precise or coarse) that would "help the investigator [who lost the original transcript] reproduce the facts of the case."

The main results of this experiment are reproduced in Fig. 8. With regard to Phase 1 performance (solid lines), the accuracy of the participants' answers declined significantly over the one-week retention interval, particularly in the first 24 h (between immediate and one-day testing). Consistent with previous findings in the (recognition) literature, the rate of decline was somewhat more shallow for the coarse answers than for the fine-grained (precise) answers, with coarse-grain accuracy substantially higher than precise-grain accuracy across all three retention intervals. More interesting are the results from the second phase (dotted line), in which participants could choose which grain size to provide for each item. As predicted, the tendency to prefer the coarse-grained answer increased with retention interval (from 43% at immediate

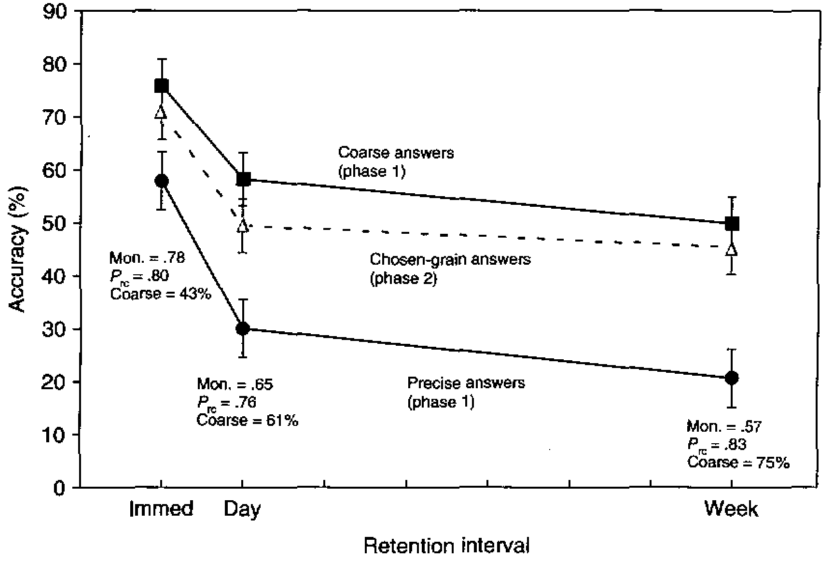

Fig. 8. Forgetting curves showing actual memory accuracy performance (mean percent correct) as a function of retention interval for the participants in Goldsmith et al. (2005, Experiment 1) plotted separately for precise-grain answers (Test phase 1), coarse-grain answers (Test phase 1), and "chosen-grain" answers (Test phase 2) for which grain size was under the control of the participant. Three further indexes are presented for each retention interval: Mon. = monitoring effectiveness, in terms of mean gamma correlation between confidence in the precise answer and its actual correctness; P_{rc} = mean grain-report criterion, estimated by examining the relationship between confidence in the precise answer and grain choice; Coarse = percentage of answers chosen at the coarse grain level in Phase 2 (as an index of reduction in informativeness). The error bars represent 95% confidence intervals.

testing to 75% after one week). This shift allowed rememberers to maintain a higher and more stable level of report accuracy than what they would have achieved had they been required to provide only precise answers. Of course, the increased accuracy was again "purchased" at the cost of reduced informativeness of the answers that were provided at the longer retention intervals.

Insight into the metacognitive mechanisms was gained by examining the confidence data. As in the earlier study, there was a strong relationship between confidence in the fine-grained answer and the grain-control decision (mean gamma = .85, across the retention intervals). Moreover, a simple satisficing model was again supported by the data: Multiple regression analyses that included both confidence in the fine-grained answer and confidence in the coarse-grained answer as predictors of the grain decision, yielded no added contribution of confidence in the coarse-grained answer, beyond what could be accounted for by confidence in the fine-grained answer alone.

Interestingly, the estimated report criterion set by the participants was equivalent at the three retention intervals, averaging around .80. This suggests that the participants were aiming to achieve approximately the same level of accuracy (80% or higher), regardless of the retention interval. Yet, the level of accuracy actually achieved was much lower than 80% (particularly at delayed testing), and accuracy did not remain stable between immediate and one-day testing. This discrepancy appears to stem from the participants' imperfect level of monitoring effectiveness, and from a decline in the level of that effectiveness over time which mirrors the pattern of decline in report accuracy: Overconfidence, in terms of the difference between mean assessed probability correct of the fine-grained answers and actual proportion correct, averaged .16 at immediate testing and .32 at one-day and one-week testing. A similar pattern of decrement over time was found for monitoring resolution (see Fig. 8). At the same time, the stability of accuracy between one-day and one-week testing appears to derive from (a) the adoption of a constant grain-control policy (report criterion) across this interval, and (b) the stability of monitoring effectiveness (in terms of both calibration and resolution) across this interval.

These results further demonstrate the critical contribution of metacognitive monitoring and control processes to memory performance. The shift in the preferred grain size with retention interval suggests an additional means by which rememberers can compensate for their failing memory: By regulating the coarseness of their answers, rememberers can maintain relatively high levels of memory accuracy despite increased forgetting. In a similar manner, perhaps rememberers can also regulate grain size to compensate for differences in other factors that affect memory, such as viewing conditions or being questioned about central versus peripheral details. Another important factor to be examined is the regulation of grain size in old age: In light of the general finding of increased reliance on gist memory in old age (e.g., Earles, Kersten, Turner, & McMullen, 1999; Koutstaal & Schacter, 1997), control over grain size in memory reporting could be a potent tool used by older rememberers to maintain their report accuracy in the face of declining memory for details.

The implications of control over grain size are, of course, especially pertinent to free-narrative and other types of open-ended memory testing procedures commonly used in naturalistic memory research. In fact, with regard to the effects of retention interval, it is remarkable that some of the studies that used such procedures (but not all of them) observed very high and stable levels of accuracy over retention intervals of up to several years (e.g., Ebbesen & Rienick, 1998; Flin, Boon, Knox, & Bull, 1992; Hudson & Fivush, 1991; Poole & White, 1991, 1993)! These levels of accuracy may have been achieved because the free-narrative format of memory report allowed participants both control over report option—withholding information that they are not sure about—as well as control over the grain size—choosing the level of precision or coarseness of the information that they reported.

V. Toward an Integrated Model of Grain Size and Report Option

Indeed, in most real-life memory situations, rememberers have the freedom both to withhold particular items of information and to choose an appropriate grain size for the information that they do report. In the research described so far, we addressed each of these two types of control separately. Now, however, returning to our hypothetical courtroom witness, we ask, how would she manage the utilization of both types of control simultaneously? When would she choose to provide a coarse-grained answer and when would she choose to respond "don't know"? In this section, we present work in progress that points toward some preliminary answers to these questions.

A. An Integrated Satisficing Model

We begin with the observation that report option and grain size may be viewed as a continuum: Withholding an answer is informationally equivalent to providing an extremely coarse-grained response that encompasses the entire range of possible values, so that it conveys no information at all about the solicited value. A simple model that builds on this idea is sketched in Fig. 9 (solid lines only). In this model, which is essentially a generalization of the satisficing model of control over grain size discussed earlier, the rememberer generates candidate answers to each question at various grain sizes, providing the most precise (informative) answer that passes a preset

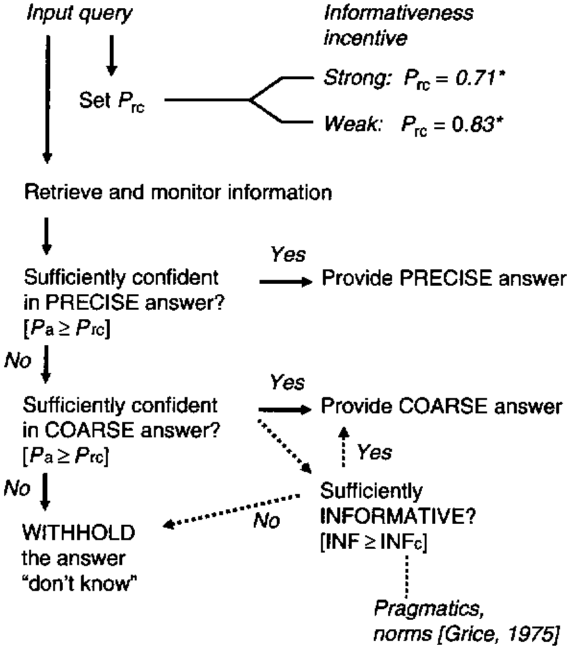
Input query
Set P_{rc}
Informativeness incentive
Strong: $P_{rc} = 0.71^*$
Weak: $P_{rc} = 0.83^*$
Retrieve and monitor information
Sufficiently confident in PRECISE answer? [$Pa \geq Prc$]
Yes
Provide PRECISE answer
No
Sufficiently confident in COARSE answer? [$Pa \geq Prc$]
Yes
Provide COARSE answer
No
WITHHOLD the answer "don't know"
No
Sufficiently INFORMATIVE? [$INF \geq INFc$]
Yes
Pragmatics, norms [Grice, 1975]

Fig. 9. Schematic outline of an integrated model of joint control over report option and grain size. See text for explanation. The P_{rc} (report criterion) values designated by an asterisk are mean empirical P_{rc} values obtained in unpublished new data. The dashed arrows designate a modification to our original model, which adds a minimum informativeness criterion (INF_c) that must be passed in order that an answer is volunteered. INF, subjectively evaluated informativeness of the answer at a particular grain size.

confidence criterion. If even the most coarsely grained candidate fails to pass the criterion, the answer is withheld entirely.

In an experiment designed to test this model, we used the same basic procedure as in our earlier semantic-memory grain-size study (Goldsmith et al., 2002), with the change that now in the second phase, the participants were allowed either to provide an answer at one of the two grain sizes, or to withhold the answer entirely. In analyzing the relationship between confidence in the answers and the choice of response (fine answer, coarse answer, or "don't know"), we found that participants used exactly the same report criterion for the control of grain size as they did for the report-option decision: If confidence in the fine-grained answer was less than .83 (mean estimated report criterion for control over grain size), participants preferred to provide the coarse-grained answer. If, however, confidence in the coarse-grained answer was also less than .83 (mean estimated report criterion for report option), the answer would be withheld entirely. Interestingly, this pattern repeated itself in another condition in which the incentive for informativeness versus accuracy was increased by increasing the monetary bonus for correct answers at the fine grain size (as in Goldsmith et al., 2002, Experiment 3). A more liberal report criterion was adopted (.71), which again was identical for the decision whether to provide the fine-grained or coarse-grained answer, and for the decision whether to provide the coarse-grained answer or instead to withhold the answer entirely.

B. To Coarsen or Withhold?

Although these data provide very nice support for the integrated satisficing model depicted in Fig. 9 [but see Weber & Brewer (in press) who obtained somewhat different estimates for the grain-choice and report-option criteria], admittedly, both the model and the experimental procedure are oversimplified: Clearly, in real-life control of grain size, rememberers are not confined to just two possible grain sizes. Instead, in principle, they have unlimited control over the grain size of their answers,[9] and hence can choose to provide as coarse an answer as is needed to reach the desired level of confidence. Why, then, would rememberers ever choose to utilize the "don't know" option under such conditions? For example, if participants could be 90% sure that Neil Armstrong walked on the moon sometime between the year 1950 and 2007, would they not prefer to provide that answer, rather than respond "don't know"?

To examine this question, we ran a further experiment using the episodic memory materials and questions from our earlier study of grain control over time (Goldsmith et al., 2005), but this time participants could write down an interval of any size as their preferred answer, or they could respond "don't know." Thus, they were allowed complete control over the grain size of their answers (see also Goldsmith et al., 2005, Experiment 2), as well as the option of withholding the answer entirely. Under such conditions, the participants chose to utilize the "don't know" option for an average of 18% of their responses (13% on immediate testing, 17% after one day, and 24% after one week). That is, in a substantial number of cases, the participants chose to refrain from providing any information at all, even though conceivably, they

[9] We acknowledge that in general, the set of candidate grain sizes that are actually considered may conform to natural linguistic units (e.g., year, decade, century) and grain sizes that cross natural linguistic boundaries may not be considered seriously by the rememberer (e.g., "between 1961 and 1971" would generally be a less natural response compared to "between 1960 and 1970" or "sometime in the 1960s"). This is an additional complication that deserves examination in future research.

could have provided at least some information in a very coarse answer that was likely to be correct (e.g., "Benny drank between 0 and 15 beers).

This interpretation could be wrong, however. Perhaps the participants chose to withhold answers only in those cases in which they could not provide any "information" even by choosing a very coarse answer. This might happen, for example, if in order to reach a reasonably high level of confidence, they would have to coarsen their answer so much that it would no longer yield any reduction in uncertainty about the solicited value beyond what could be inferred on the basis of general knowledge of the world (e.g., script knowledge) and common sense alone. To examine this possibility, we had the participants go back to each of their "don't know" answers and now provide an interval answer that they were 100% sure was correct. When these 100% confidence intervals (CIs) were compared to those given by a group of control participants who read the stimulus story with all of the target quantitative information blacked out (preventing any episodic memory of the actual quantities), the CIs of the experimental participants were about half as wide[10] as those provided by the unexposed control participants, even though the experimental participants had responded "don't know" to these questions initially.

These results indicate that the experimental participants chose the "don't know" option even though they could have provided some useful information to an outsider (e.g., a police investigator) who had knowledge only of the general episode and not of the actual quantities. Moreover, the answer provided by the participants in the second phase was not only subjectively informative, it was also objectively informative: The 100% CIs obtained for the "don't know" items of the experimental participants were significantly more likely to contain the correct target value than were the 100% CIs obtained from a second group of control participants, who answered the same questions at the same interval widths used by the experimental participants, on the basis of common sense and script knowledge alone.

C. The Need for an Informativeness Criterion

How do the preceding results bear on the idea of an integrated model of grain size and report option? It appears that what was lacking in the original satisficing model of grain control, and would be needed in a new integrated model, is a minimum *informativeness criterion* that must be satisfied by any reported answer, in addition to the minimum confidence (accuracy) criterion

[10] In order to approximately equate the contribution of different items to the overall interval-width differences, the interval widths were normalized by the midpoint of the answers (normalized interval width = actual interval width/interval midpoint).

that we have been focusing on exclusively in our work so far. This addition to the basic model is depicted schematically by the dashed lines in Fig. 9.

According to social and pragmatic norms of communication, people are expected not only to be reasonably accurate in what they report, but also to be reasonably informative (Grice, 1975). We assume, then, that very coarse-grained answers, such as "Benny drank between 0 and 15 beers," are generally avoided because they violate these norms of communication. These norms, together with more specific contextual considerations, presumably affect the setting of the minimum informativeness criterion. If one's knowledge or memory is so poor that one can only reach a high level of confidence by providing such an answer, the normatively acceptable option (and one that does not violate the minimum informativeness criterion) is to refrain from providing an answer, responding instead, "don't know."

Note that by this analysis, rememberers should make use of the "don't know" option specifically when they feel that they are unable to provide an answer that is both sufficiently accurate (likely to be correct) and sufficiently informative. One can imagine real-life situations, however, in which there are implicit or explicit demands to provide a substantive answer (i.e., "don't know" is not permitted). What should rememberers do in such situations when they find themselves unable to simultaneously satisfy the confidence and informativeness criteria?

As part of a doctoral research project, currently underway, Ackerman and Goldsmith (2006) put forward a distinction between two knowledge states: *satisficing* and *unsatisficing* knowledge. Whereas in the former state one has sufficient knowledge (or memory) to support an answer that simultaneously satisfies the confidence and informativeness criteria, in the latter state one does not. They proposed that in a state of unsatisficing knowledge, withholding of answers ("don't know" response) is the preferred way of resolving the criterion conflict, but when this option is denied, rememberers have no choice but to violate one or both of the two criteria. In that case, they should tend to sacrifice the confidence criterion rather than the informativeness criterion. This is because the penalty (e.g., ridicule) for providing an overly uninformative answer is often immediate, whereas the accuracy or inaccuracy of one's answer is generally only evident at a later time, if at all. The results of several experiments are consistent with these ideas, yielding the following general conclusions: (1) Participants strive to satisfy a minimum informativeness criterion as well as a minimum confidence-accuracy criterion. (2) Criterion conflicts are more likely to occur when knowledge (memory) is low. (3) When criterion conflicts occur, participants will often violate the confidence criterion in order to meet the informativeness criterion. (4) Given joint control over grain size and report option, participants tend to circumvent the criterion conflict when it occurs, by withholding the answers entirely.

VI. Conclusion

The work described in this chapter is predicated on the view that the strategic regulation of memory performance is an intrinsic aspect of everyday remembering. Therefore, to achieve a more complete understanding of remembering in real-life contexts, it is important to identify the various types of control that people exert over their memory reporting, and examine their underlying mechanisms and performance consequences. The desire to capture the full richness of real-world memory phenomena, however, is often at odds with the desire to bring the phenomena into the laboratory for controlled experimental investigation. In our work, we have tried to reach an expedient compromise that offers the benefits of experimental tools and rigor while still tapping some of the fundamental features of the strategic regulation of memory performance in real-world settings.

Acknowledgments

The preparation of this chapter was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of German-Israeli Project Cooperation (DIP). We thank the BMBF for their support.

References

Abu-Sayf, F. K. (1979). The scoring of multiple-choice tests: A closer look. *Educational Technology, 19,* 5–15.

Ackerman, R., & Goldsmith, M. (2006). *Control over grain size in question answering with unsatisficing knowledge.* Poster presented at the affect, motivation, and decision-making International Conference, Ein Boqeq, The Dead Sea, Israel.

Albanese, M. A. (1988). The projected impact of the correction for guessing on individual scores. *Journal of Educational Measurement, 25,* 149–157.

American Psychologist. (1991). 46 (1).

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.

Angoff, W. H. (1989). Does guessing really help? *Journal of Educational Measurement, 26,* 323–336.

Angoff, W. H., & Schrader, B. W. (1984). A study of hypotheses basic to the use of rights and formula scores. *Journal of Educational Measurement, 21,* 1–17.

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 integrative system of metamemory components involved in retrieval. In D. Gopher and A. Koriat (Eds.) *Cognitive regulation of performance: Interaction of theory and application. Attention and Performance XVII* (pp. 287–313). Cambridge, MA: MIT Press.

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

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

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*, 55–68.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. P. Shimamura (Eds.) *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

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.

Brewer, W. F. (1992). Phenomenal experience in laboratory and autobiographical memory tasks. In M. A. Conway, D. C. Rubin, H. Spinnler, and W. A. Wagenaar (Eds.) *Theoretical perspectives on autobiographical memory* (pp. 31–51). Dordrecht: Kluwer.

Brown, E. L., Deffenbacher, K. A., & Sturgill, W. (1977). Memory for faces and the circumstances of encounter. *Journal of Applied Psychology, 62*, 311–318.

Brown, J. (Ed.). (1992). *Recall and recognition.* London: Wiley.

Bruck, M., & Ceci, S. J. (1999). The suggestibility of children's memory. *Annual Review of Psychology, 50*, 419–439.

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.

Burgess, P. W., & Shallice, T. (1996). Confabulation and the control of recollection. *Memory, 4*, 359–411.

Butler, K. M., McDaniel, M. A., Dornburg, C. C., Price, A. A., & Roediger, H. L., 3rd (2004). Age differences in veridical and false recall are not inevitable: The role of frontal lobe function. *Psychonomic Bulletin & Review, 11*, 921–925.

Cassel, W. S., Roebers, C. M., & Bjorklund, D. F. (1996). Developmental patterns of eyewitness responses to repeated and increasingly suggestive questions. *Journal of Experimental Child Psychology, 61*, 116–133.

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

Conway, M. A. (1995). *Flashbulb memories.* Hove, UK: Erlbaum.

Conway, M. A., Collins, A. F., Gathercole, S. E., & Anderson, S. J. (1996). Recollections of true and false autobiographical memories. *Journal of Experimental Psychology: General, 125*, 69–95.

Cronbach, L. J. (1984). *Essentials of psychological testing.* New York: Harper & Row.

Cross, L. H., & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. *Journal of Educational Measurement, 14*, 313–321.

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.

Donaldson, W. (1992). Measuring recognition memory. *Journal of General Psychology, 121*, 275–277.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment. *Psychological Science, 5*, 69–106.

Earles, J. L., Kersten, A. W., Turner, J. M., & McMullen, J. (1999). Influences of age, performance, and item relatedness on verbatim and gist recall of verb-noun pairs. *Journal of General Psychology, 126*, 97–110.

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

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

Erdelyi, M. H., & Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials. *Cognitive Psychology, 6,* 159–171.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552–564.

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

Fisher, R. P., & Craik, F. I. M. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 701–711.

Fisher, R. P., Geiselman, R. E., & Raymond, D. S. (1987). Critical analysis of police interview techniques. *Journal of Police Science and Administration, 15,* 177–185.

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.

Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement, 4,* 79–90.

Fruzzetti, A. E., Toland, K., Teller, S. A., & Loftus, E. F. (1992). Memory and eyewitness testimony. In M. M. Gruneberg and P. E. Morris (Eds.) *Aspects of memory* (Vol. 1, pp. 18–50). London, UK: Routledge.

Gafni, N. (1990). *Differential tendencies to guess as a function of gender and lingual-cultural reference group* (Report No. 115). Jerusalem, Israel: National Institute for Testing and Evaluation.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10,* 843–876.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

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

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

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

Goodman, G. A. (2006). Children's eyewitness memory: A modern history and contemporary commentary. *Journal of Social Issues, 62,* 811–832.

Gorenstein, G. W., & Ellsworth, P. C. (1980). Effect of choosing an incorrect photograph on a later identification by an eyewitness. *Journal of Applied Psychology, 65,* 616–622.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory andpsychophysics.* New York: Wiley.

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

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75,* 424–429.

Healy, A. F., & Jones, C. (1973). Criterion shifts in recall. *Psychological Bulletin, 79,* 335–340.

Hertzog, C., & Dunlosky, J. (2004). Aging, metacognition, and cognitive control. In B. H. Ross (Ed.) *The psychology of learning and motivation: Advances in research and theory* (pp. 215–247). San Diego, CA: Elsevier.

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.

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

Higham, P. A., & Gerrard, C. (2005). Not all error are created equal: Metacogntion and changing answers on multiple-choice tests. *Canadian Journal of Experimental Psychology, 59*, 28–34.

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

Higham, P. A., & Tam, H. (2006). Release from generation failure: The role of study-list structure. *Memory & Cognition, 34*, 148–157.

Hilgard, E. R., & Loftus, E. F. (1979). Effective interrogation of the eyewitness. *International Journal of Clinical and Experimental Hypnosis, 27*, 342–357.

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

Jacoby, L. L. (1999). Deceiving the elderly: Effects of accessibility bias in cued-recall performance. *Cognitive Neuropsychology, 16*, 417–436.

Jacoby, L. L., & Rhodes, M. G. (2006). False remembering in the aged. *Current Directions in Psychological Science, 15*, 49–53.

Jacoby, L. L., Debner, J. A., & Hay, J. F. (2001). Proactive interference, accessibility bias, and process dissociations: Valid subject reports of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 686–700.

Janowsky, J. S., Shimamura, A. P., & Squire, L. R. (1989). Memory and metamemory: Comparisons between frontal lobe lesions and amnesic patients. *Psychobiology, 17*, 3–11.

Johnson, M. K. (1997). Identifying the origin of mental experience. In M. S. Myslobodsky (Ed.) *The Mythomanias: The nature of deception and self deception* (pp. 133–180). Mahwah, NJ: Erlbaum.

Kato, T. (1985). Semantic-memory sources of episodic retrieval failure. *Memory & Cognition, 13*, 442–452.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35*, 157–175.

Kelley, C. M., & Jacoby, L. L. (1998). Subjective reports and process dissociation: Fluency, knowing, and feeling. *Acta Psychologica, 98*, 127–140.

Kelley, C. M., & Jacoby, L. L. (2000). Recollection and familiarity. In E. Tulving and F. I. M. Craik (Eds.) *The Oxford handbook of memory* (pp. 215–228). London: Oxford University Press.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*, 1–24.

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

Klatzky, R. L., & Erdelyi, M. H. (1985). The response criterion problem in tests of hypnosis and memory. *International Journal of Clinical and Experimental Hypnosis, 33*, 246–257.

Koren, D., Poyurovsky, M., Seidman, L. J., Goldsmith, M., Wenger, S., & Klein, E. (2005). The neuropsychological basis of competence to consent in first-episode schizophrenia: A pilot metacognitive study. *Biological Psychiatry, 57*, 609–616.

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

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

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*, 311–333.

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, and E. Thompson (Eds.) *Cambridge handbook of consciousness* (pp. 289–325). Cambridge, UK: Cambridge University Press.

Koriat, A., & Goldsmith, M. (1997). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General, 123*, 297–316.

Koriat, A., & Goldsmith, M. (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 and 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. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken and Y. Trope (Eds.) *Dual process theories in social psychology* (pp. 483–502). New York: Guilford Press.

Koriat, A., Ben-zur, H., & Sheffer, D. (1988). Telling the same story twice: Output monitoring and age. *Journal of Memory and Language, 27*, 23–39.

Koriat, A., Goldsmith, M., & Halamish, V. (in press). Control processes in voluntary remembering. In H. L. Roediger, III (Ed.), *Cognitive psychology of memory*. Vol. 2 of *Learning and memory: A comprehensive reference*, 4 vols. (J. Byrne, Editor). Oxford, UK: Elsevier.

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

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.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*, 36–69.

Koutstaal, W. (2006). Flexible remembering. *Psychonomic Bulletin & Review, 13*, 84–91.

Koutstaal, W., & Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language, 37*, 555–583.

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

Lipton, J. P. (1977). On the psychology of eyewitness testimony. *Journal of Applied Psychology, 62*, 90–95.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74*, 100–109.

Massin-Krauss, M., Bacon, E., & Danion, J.-M. (2002). Effects of the benzodiazepine lorazepam on monitoring and control processes in semantic memory. *Consciousness and Cognition, 11*, 123–137.

Memon, A., & Stevenage, V. S. (1996). Interviewing witnesses: What works and what doesn't? *Psycholoquy, 7*, witness-memory.1.memon.

Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing*. Cambridge: MIT Press.

Mitchell, K. J., & Johnson, M. K. (2000). Source monitoring: Attributing mental experiences. In E. Tulving and F. I. M. Craik (Eds.) *The Oxford handbook of memory* (pp. 179–195). London: Oxford University Press.

Moritz, S., & Woodward, T. S. (2006). The contribution of metamemory deficits to schizophrenia. *Journal of Abnormal Psychology, 15*, 15–25.

Moritz, S., Woodward, T. S., & Chen, E. (2006). Investigation of metamemory dysfunctions in first-episode schizophrenia. *Schizophrenia Research, 81*, 247–252.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519–533.

Moscovitch, M., & Winocur, G. (1995). Frontal lobe, memory, and aging. In J. Grafman, J. K. Holyoak, and F. Boller (Eds.) *Structure and functions of the human prefrontal cortex* (Vol. 769, pp. 119–150). New York, NY: New York Academy of Science.

Moston, S. (1987). The suggestibility of children in interview studies. *First Language, 7*, 67–78.

Mulder, M. R., & Vrij, A. (1996). Explaining conversation rules to children: An intervention study to facilitate children's accurate responses. *Child Abuse and Neglect, 20*, 623–631.

Neisser, U. (1988). Time present and time past. In M. M. Gruneberg, P. Morris, and R. Sykes (Eds.) *Practical aspects of memory: Current research and issues* (Vol. 2, pp. 545–560). Chichester, England: 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. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. *Applied Cognitive Psychology, 10*, 257–260.

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

Nilsson, L.-G. (1987). Motivated memory: Dissociation between performance data and subjective reports. *Psychological Research, 49*, 183–188.

Norman, D. A., & Wickelgren, W. A. (1969). Strength theory of decision rules and latency in retrieval from short-term memory. *Journal of Mathematical Psychology, 6*, 192–208.

Norman, K. A., & Schacter, D. L. (1996). Implicit memory, explicit memory, and false recollection: A neuroscience perspective. In L. M. Reder (Ed.) *Implicit memory and metacognition* (pp. 229–257). Hillsdale, NJ: Erlbaum.

Notea-Koren, E. (2006). *Performance accuracy and quantity in psychometric testing: An examination and assessment of cognitive and metacognitive components*. Unpublished doctoral dissertation, University of Haifa, Israel.

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

Pansky, A., Koriat, A., Goldsmith, M., & Pearlman, S. (2002). *Memory accuracy and distortion in old age: Cognitive, metacognitive, and neurocognitive determinants*. Poster presented at the

30th Anniversary Conference of the National Institute for Psychobiology in Israel, Hebrew University, Jerusalem, Israel.

Parks, T. E. (1966). Signal detectability theory of recognition-memory performance. *Psychological Review, 73,* 44–58.

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.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 435–451.

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.

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

Roebers, C. M., & Fernandez, O. (2002). The effects of accuracy motivation and children's and adults' event recall, suggestibility, and their answers to unanswerable questions. *Journal of Cognition and Development, 3,* 415–443.

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

Roediger, H. L. (1980). Memory metaphors in cognitive psychology. *Memory & Cognition, 8,* 231–246.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory & Cognition, 21,* 803–814.

Rosenbluth-Mor, M. (2001). *Accuracy and quantity in memory reports: The effects of context reinstatement.* Unpublished master's thesis, University of Haifa. Israel.

Ross, M. (1997). Validating memories. In N. L. Stein, B. Ornstein, B. Tversky, and C. Brainerd (Eds.) *Memory for everyday and emotional events* (pp. 49–81). Mahwah, NJ: Erlbaum.

Schacter, D. (1990). Memory. In M. I. Posner (Ed.) *Foundation of cognitive science* (pp. 687–725). Cambridge, MA: MIT Press.

Schacter, D. L., Norman, D. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology, 49,* 289–318.

Schacter, D. L., Verfaellie, M., & Pradere, D. (1996). The neuropsychology of memory illusions: False recall and recognition in amnesic patients. *Journal of Memory and Language, 35,* 319–334.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feeling of knowing. *Psychonomic Bulletin & Review, 1,* 357–375.

Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1074–1083.

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

Slakter, M. J. (1968). The penalty for not guessing. *Journal of Educational Measurement, 5,* 141–144.

Son, L. K., & Schwartz, B. L. (2002). The relation between metacognitive monitoring and control. In T. J. Perfect and B. S. Schwartz (Eds.) *Applied metacognition* (pp. 15–38). Cambridge, UK: Cambridge University Press.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68*, 301–340.

Thiede, K. W., Anderson, C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66–73.

Thomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology, 86*, 255–262.

Thurstone, L. L. (1919). A method for scoring tests. *Psychological Bulletin, 16*, 235–240.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: The Clarendon Press.

Tulving, E., & Osler, S. (1968). Effectiveness of retrieval cues in memory for words. *Journal of Experimental Psychology, 77*, 593–601.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 352–373.

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

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 582–600.

Weber, N., & Brewer, N. (in press). Eyewitness recall: Regulation of grain size and the role of confidence. *Journal of Experimental Psychology: Applied.*

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

Winograd, E. (1996). Contexts and functions of retrieval. *Behavioral and Brain Sciences, 19*, 209–210.

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review, 107*, 369–376.

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. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*, 611–617.

Zaragoza, M. S., & Mitchell, K. J. (1996). Repeated exposure to suggestion and the creation of false memories. *Psychological Science, 7*, 294–300.

Zeelenberg, R. (2005). Encoding specificity manipulations do affect *retrieval* from memory. *Acta Psychologica, 119*, 107–121.