Psychological Review
1996, Vol. 103, No. 3, 490-517

Monitoring and Control Processes in the Strategic Regulation of Memory Accuracy

Asher Koriat and Morris Goldsmith
University of Haifa

When people are allowed freedom to volunteer or withhold information, they can enhance the accuracy of their memory reports substantially relative to forced-report performance. A theoretical framework addressing the strategic regulation of memory reporting is put forward that delineates the mediating role of metamemorial monitoring and control processes. Although the enhancement of memory accuracy is generally accompanied by a reduction in memory quantity, experimental and simulation results indicate that both of these effects depend critically on (a) accuracy incentive and (b) monitoring effectiveness. The results are discussed with regard to the contribution of metamemory processes to memory performance, and a general methodology is proposed that incorporates these processes into the assessment of memory-accuracy and memory-quantity performance.

"Write down as many words as you can remember from the list that was presented to you earlier." These are the free-recall instructions that a participant typically receives in a list-learning experiment. These instructions explicitly define a clear goal for the participant: to retrieve the largest possible number of studied items. Other requirements, however, are often left unspecified, and participants' responses may be guided by implicit assumptions about these requirements. For example, participants will generally terminate their memory search at some point when they judge that the set of accessible words has been depleted or that more effort would be futile (Barnes, Nelson, Dunlosky, Mazzoni, & Narens, 1995; Costermans, Lories, & Ansay, 1992; Gruneberg, Monks, & Sykes, 1977; Nelson et al., 1990; Nelson & Narens, 1990). Participants also generally assume that they are not to repeat a word that has already been reported and therefore tend to monitor each retrieved word for its previous occurrence in the output list before they report it (Gardiner & Klee, 1976; Gardiner, Passmore, Herriot, & Klee, 1977; Koriat, Ben-Zur, & Sheffer, 1988; Murdock, 1974). Of more direct relevance to our present concerns, even when there are no explicit instructions about guessing, participants may be guided by the assumption that they are expected not only to reproduce a large proportion of the studied words but also to be accurate. Indeed, when participants are specifically instructed to be less "inhibited," additional items may be reported, but the majority of these are commission errors (e.g., Bousfield & Rosner, 1970; Erdelyi, Finks, & Feigin-Pfau, 1989; Roediger & Payne, 1985). Thus, even in a simple free-recall testing situation, memory performance probably reflects a variety of metamemorial monitoring and control processes that help participants achieve both explicit and implicit performance goals.

The strategic control of memory reporting undoubtedly plays an even greater role outside the laboratory. In the many real-life situations in which people recount past events, numerous types of decisions are routinely made about what material to report and in what amount of detail (e.g., Fisher, Geiselman, & Raymond, 1987; Flanagan, 1981; Hilgard & Loftus, 1979; Neisser, 1981, 1988; Neisser & Fivush, 1994; Nigro & Neisser, 1983).

In this article, we focus on one way in which people regulate their memory reporting in the service of memory accuracy—by deciding which items of information to volunteer and which to withhold. Although, as just mentioned, this type of regulation probably occurs in many laboratory situations, it certainly plays a prominent role in many real-life memory situations as well. Consider, for example, a person on the witness stand who has sworn to "tell the truth, the whole truth, and nothing but the truth." Compared with the typical list-learning instructions illustrated earlier, the instructions to the witness set a more ambitious goal: to provide a memory report that is both complete and entirely accurate. In striving to meet that goal, the witness presumably must weigh the risks ensuing from commission and omission errors before deciding either to answer a given question or else to respond "I don't remember."[1]

In addressing the mechanisms underlying the strategic regulation of memory accuracy, we put forward a theoretical framework that delineates the mediating role of monitoring and control processes. The framework is used to investigate the manner

Asher Koriat and Morris Goldsmith, Department of Psychology, University of Haifa, Haifa, Israel.

This research was supported by Grant 032-1106 (1990) from the Israel Foundations Trustees and was conducted while Morris Goldsmith was supported by a National Science Foundation Graduate Fellowship. The work was carried out at the Institute of Information Processing and Decision Making, University of Haifa. We thank Ian Begg, Ron Fisher, Larry Jacoby, and Colleen Kelley for their helpful comments on a previous version of the article. We also thank Michal Sion for her help in setting up the experiments and in the preparation of the manuscript.

Correspondence concerning this article should be addressed to Asher Koriat, Department of Psychology, University of Haifa, Haifa 31905, Israel. Electronic mail may be sent via Internet to akoriat@psy.haifa.ac.il.

[1] Another prominent means of regulation in such situations, which will not be addressed here, is control over the "grain size" or level of generality of the reported answer (see Fisher, 1996; Koriat & Goldsmith, 1996a, 1996c; Neisser, 1988; Yaniv & Foster, 1995, in press).

in which these processes affect both the accuracy and the quantity of reported information. This investigation is motivated by our previous work, which suggests that subject-controlled regulatory processes may play a crucial role in enabling accurate memory reporting in real-life memory situations. We begin, then, by summarizing that work, which provides the basis for the present theoretical development.

Review of Previous Work

When considering the contribution of subject-controlled processes to memory performance, it is critical to distinguish between two different properties of memory—quantity and accuracy (Klatzky & Erdelyi, 1985; Stern, 1904). As we have previously shown (Koriat & Goldsmith, 1994, 1996b, 1996c), these two properties, as well as subject control, have received rather different emphases in current research practices: Whereas traditional memory research has evaluated memory primarily in terms of the quantity of items recovered under tightly (experimenter-) controlled conditions (Banaji & Crowder, 1989; Schacter, 1989), the new wave of naturalistic, "everyday memory" research (Foss, 1991) has exhibited a greater concern for the accuracy or faithfulness of memory, coupled with a tendency to allow participants greater control over their memory reporting (Hilgard & Loftus, 1979; Neisser, 1988). Our recent work (Koriat & Goldsmith, 1994, 1996b, 1996c) indicates that some of the apparent inconsistencies that emerge when comparing findings across naturalistic and laboratory contexts may be due to the different roles played by subject control in accuracy-based and quantity-based memory assessment. To lay the ground for the present article, we shall briefly outline several points that have emerged from that work:

1. The quantity-oriented and accuracy-oriented approaches to memory reflect two fundamentally different ways of thinking about memory.

We have argued that the focus on memory quantity versus memory accuracy actually reflects a distinction between two alternative memory metaphors, the *storehouse* and the *correspondence* metaphors, respectively (Koriat & Goldsmith, 1996c). Each of these has unique implications for the study and assessment of memory. What is important in the present context is that the correspondence metaphor entails the evaluation of memory in terms of its faithfulness in representing past events, rather than merely in terms of the number of items remaining in store (Bartlett, 1932; Ross, in press; Schacter, 1989, 1995; Winograd, 1994). The emergence of this metaphor can be seen in the growing body of work on eyewitness testimony, autobiographical memory, metamemory, memory distortions, false memories, and other memory phenomena in which the reliability or unreliability of memory is of primary concern (e.g., Brewer, 1996; Loftus, 1979, 1982; Neisser, 1981, 1988; Neisser & Fivush, 1994; Ross, in press; Schacter, Coyle, Fishbach, Mesulam, & Sullivan, 1995; Winograd & Neisser, 1992). Indeed, inspection of such work shows the correspondence-oriented evaluation of memory accuracy to have its own unique logic, which departs from that underlying the traditional quantity-oriented approach to memory (Koriat & Goldsmith, 1996c).

2. Quantity-based memory measures are inherently input-bound, whereas accuracy-based measures are output-bound.

There are cases in which accuracy-based memory measures can be qualitatively different from the traditional quantity-based measures (Koriat & Goldsmith, 1996b, 1996c). In the context of the more standard, item-based approach to memory (Puff, 1982), however, memory quantity and accuracy measures are distinguished in terms of the difference between *input-bound* and *output-bound* assessment (Koriat & Goldsmith, 1994): Quantity measures, traditionally used to tap the amount of studied information that can be recovered, are input-bound, reflecting the likelihood that each *input* item is correctly remembered (e.g., the percentage of studied items recalled or recognized). Accuracy measures, in contrast, evaluate the dependability of memory—the extent to which remembered information can be trusted to be correct. Hence, these measures are output bound: They reflect the conditional probability that each *reported* item is correct.

The output-bound assessment of memory accuracy is particularly suited to situations such as eyewitness testimony, in which a high premium is placed on obtaining memory reports that can be relied on (see, e.g., Deffenbacher, 1991; Fisher, Geiselman, & Amador, 1989; Hilgard & Loftus, 1979; Loftus, 1979; Poole & White, 1991, 1993; Wells & Lindsay, 1985; Wells & Loftus, 1984). Thus, for instance, a witness might recall the names of three out of five people present at some target event, achieving only 60% quantity. Yet, if those three people were indeed present, then failing to include the other two names would not detract from the accuracy of the information that was reported (100% accuracy). On the other hand, naming a fourth person who was not present would reduce the accuracy (but not the quantity) of the report (to 75%). Essentially, then, whereas input-bound measures hold the rememberer responsible for what he or she fails to report, output-bound measures hold the rememberer accountable only for what he or she does report.

3. Input-bound (quantity) and output-bound (accuracy) measures are operationally distinguishable only under free-report conditions.

Despite the different emphases of input-bound and output-bound memory assessment, there are conditions in which the two types of measures are operationally equivalent: The critical factor is *report option,* that is, whether or not the rememberer is required to answer all items. When memory is tested through a forced-report procedure, memory quantity and accuracy measures are necessarily equivalent, because the likelihood of remembering each input item (quantity) is equal to the likelihood that each reported item is correct (accuracy). Under such conditions, the distinction between accuracy-based and quantity-based memory assessment is solely a matter of the experimenter's intent (e.g., Loftus, Miller & Burns, 1978; see Koriat & Goldsmith, 1994, 1996c).

Accuracy and quantity measures can differ substantially, however, under free-report conditions, in which rememberers are implicitly or explicitly given the option either to volunteer a piece of information or to abstain (e.g., respond "I don't know"; Neisser, 1988). Most everyday situations are of this sort. In the laboratory, the most typical example is the standard free-recall task, in which reporting is essentially controlled by the participant. Under free-report conditions, people tend to provide only information that they believe is likely to be correct, so that their performance is mediated by a decision process used to avoid incorrect answers (Klatzky & Erdelyi, 1985; Koriat & Goldsmith, 1994). Because the number of volunteered answers is generally smaller than the number of input items, the output-bound (accuracy) and input-bound (quantity) memory measures can vary substantially. Thus, there is an intrinsic connection between the experimental focus on output-bound accuracy, on the one hand, and the provision of subject control over memory reporting, on the other (Fisher, 1996; Hilgard & Loftus, 1979; Koriat & Goldsmith, 1994, 1996b, 1996c; Neisser, 1988).

4. Report option has a substantial effect on memory accuracy performance, which can complicate the interpretation of empirical findings.

Report option is important not only because it allows memory accuracy to be operationally distinguished from memory quantity, but also because it, in itself, has a substantial effect on memory accuracy performance. This effect is generally concealed, however, because in the reality of memory research, report option tends to be confounded with other aspects of memory testing.

The contribution of report option was revealed in several experiments that addressed a seemingly discrepant pattern of results between laboratory and naturalistic research contexts (Koriat & Goldsmith, 1994). We called this pattern the *recall–recognition paradox:* On the one hand, the established wisdom in eyewitness research holds that testing procedures involving recognition or directed questioning can have "contaminating" effects on memory (see, e.g., Brown, Deffenbacher, & Sturgill, 1977; Gorenstein & Ellsworth, 1980; Hilgard & Loftus, 1979; Loftus, 1979, 1982; Loftus & Hoffman, 1989). 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 et al., 1987; Hilgard & Loftus, 1979). On the other hand, however, this body of evidence stands in seeming defiance of the well-established superiority of recognition over recall memory in traditional, list-learning laboratory experiments (e.g., Brown, 1976; Shepard, 1967; but see Tulving & Thomson, 1973). In fact, this discrepancy could be taken to suggest a difference in the dynamics of memory between naturalistic and laboratory contexts (Neisser, 1988; see also Conway, 1991, 1993; Gruneberg & Morris, 1992).

Another interpretation, however, is that because laboratory list-learning experiments have focused on memory quantity rather than accuracy, the paradox simply reflects an interaction between *test format* (recall–production vs. recognition–selection) and *memory property:* Recognition testing would be superior to recall testing in terms of memory quantity performance, but recall testing would yield better memory accuracy (see Hilgard & Loftus, 1979; Lipton, 1977; Neisser, 1988). This interpretation too, however, is complicated by the fact that testing procedures that differ in test format often differ in report option as well. For instance, in free-recall testing, people produce their own answers (production format) and report only what they feel they actually remember (free report), whereas in forced-choice recognition testing, people are not only confined to choosing between the alternatives presented by the interrogator (selection format), they are also required to answer each and every item (forced report).

To unravel the paradox and expose the effects of subject control over memory reporting, we orthogonally manipulated memory property, test format, and report option (Koriat & Goldsmith, 1994). In addition to the standard methods of free recall[2] and forced recognition, we also included the less common procedures of forced recall (in which the participants were required to answer all items) and free recognition (in which the participants were allowed to skip over items). Both quantity and accuracy scores were derived for all four methods.

In each of three laboratory experiments, the "paradoxical" pattern was obtained: Forced recognition yielded better quantity performance but poorer accuracy performance than free recall. However, when test format and report option were disentangled, it became clear that the superiority of recognition quantity performance was indeed due to test format but that the superiority of recall accuracy was entirely due to the option of free report: First, under free-report conditions, in which the recall and recognition participants had equal opportunity to screen their answers, the recognition and recall accuracy scores were virtually identical. Second, free-report accuracy performance was substantially better than forced-report performance for both the recall and the recognition test formats. Thus, even in a laboratory research context, and regardless of the particular format of the test, report option emerges as a critical factor in the assessment of memory accuracy.

5. Memory accuracy performance is under strategic control, whereas memory quantity performance is not.

The results just summarized indicate that memory accuracy performance can be improved considerably by allowing people to control their own memory reporting. Across the three experiments, the accuracy advantage of free over forced report ranged from 61% to 89% for recall and from 15% to 38% for recognition (Koriat & Goldsmith, 1994). In addition, given the option of free report, people can apparently adjust their memory accuracy in accordance with the operative level of accuracy incentive: When our free-report participants were given a very high accuracy incentive (receiving a monetary bonus for each correct answer but forfeiting all winnings if even a single incorrect answer was volunteered), they improved their accuracy performance substantially for both recall and recognition testing compared with performance under a more moderate incentive (in which the penalty for each incorrect answer equaled the bonus for each correct answer). In fact, fully one fourth of the high-incentive participants succeeded in achieving 100% accuracy! The improved accuracy, however, was accompanied by a corresponding reduction in quantity performance (i.e., in the number of correct answers provided or selected).

[2] We use the term *free recall,* in opposition to *forced recall,* to denote the option of free report. In traditional usage, however, the former term has been used in opposition to serial recall, indicating only that the participant is free to choose the order in which items are to be recalled.

These results regarding memory accuracy contrast sharply with the general observation from quantity-oriented research that people cannot improve their memory-quantity performance when given incentives to do so: First, offering recall and recognition participants monetary incentives to produce as many correct answers as possible does not increase their quantity performance relative to control participants who are not given any special incentive (e.g., Nilsson, 1987; Weiner, 1966a, 1966b; but see Loftus & Wickens, 1970). Second, studies investigating the effects of recall criterion (e.g., Bousfield & Rosner, 1970; Britton, Meyer, Hodge, & Glynn, 1980; Cofer, 1967; Erdelyi, 1970; Erdelyi et al., 1989: Keppel & Mallory, 1969; Roediger & Payne, 1985; Roediger, Srinivas, & Waddil, 1989) generally indicate that encouraging or forcing people to recall more items does not improve their memory-quantity performance relative to standard free-recall instructions (e.g., Roediger & Payne, 1985). In those experiments that did yield some improvement (e.g., Bousfield & Rosner, 1970, Experiment 5; Erdelyi et al., 1989), "large manipulations of criteria produced only small gains in correct recall" (Roediger et al., 1989, p. 256).

In sum, our work comparing the accuracy-oriented and quantity-oriented approaches to memory suggests that subject control over memory reporting may play very different roles in these two approaches: Whereas in quantity-oriented research, the effects of subject control can perhaps be disregarded on empirical grounds (Roediger et al., 1989), subject control must be taken seriously in accuracy-oriented research, because here the effects on both memory-accuracy and memory-quantity performance can be substantial. Thus, in view of the growing interest in the dependability of people's freely reported remembrances, both in the courtroom and elsewhere, more attention needs to be paid to the potential contribution of subject control under free-report memory conditions.

So far, our results suggest that when given a strong enough incentive for accuracy, people can, at least under some circumstances, provide memory reports that are very accurate. How is such accuracy achieved? Does accurate reporting necessarily come at the expense of the amount of information provided? Under what conditions can an eyewitness tell the whole truth but still report nothing but the truth? Might there be individual or group differences in the regulation of memory accuracy? These and other questions call for a more systematic effort to clarify the mechanisms by which people regulate their memory reporting. In the following section, we put forward a general theoretical framework for investigating these mechanisms and their effects on memory accuracy and quantity performance.

A Model of Free-Report Monitoring and Control

The idea that self-directed decision processes may mediate memory performance is, of course, not new. It was primarily signal-detection theory (Green & Swets, 1966; Swets, Tanner, & Birdsall, 1961) that inspired a consideration of the role of such processes in memory responding. This framework has been used extensively to investigate the decision processes underlying forced-report, recognition memory (see, e.g., Banks, 1970; Bernbach, 1967; Kintsch, 1967; Lockhart & Murdock, 1970; Murdock, 1974, 1982; Norman & Wickelgren, 1969). Thus, for instance, in the standard "old–new" recognition paradigm, participants 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 indices are typically derived: a measure of retention, d', and a measure of criterion level, β.

The situation is very different, however, with regard to the decision process underlying free-report memory performance, that is, with the decision of whether to report an answer or abstain. Unfortunately, under such conditions, the signal-detection methodology cannot be properly applied (Lockhart & Murdock, 1970). Thus, despite the fact that many memory models include a response criterion, the operation of that criterion in free-report testing situations, when considered at all, has typically been treated as a nuisance factor that should be avoided or experimentally controlled (Klatzky & Erdelyi, 1985) rather than directly investigated. Indeed, Nelson and Narens (1994) pointed out that memory researchers generally

> attempt to eliminate or reduce their subjects' variations in self-directed processing because (1) such processing on the part of the subject is typically construed mainly as a source of noise . . ., and (2) until recently, there have not been theoretical frameworks within which to systematically explore the subjects' self-directed processing. (p. 9)

As will be discussed at length later (see General Discussion), from an accuracy-oriented perspective, a research strategy that takes control away from the rememberer is unsatisfactory. Thus, in our proposed framework, we extend the basic logic underlying signal-detection theory to free-report situations (cf. Klatzky & Erdelyi, 1985) but also augment it with concepts and methods borrowed from the study of metamemory.

Metamemory research has generally focused on the capacity or ability aspects of metamemory, that is, what people "know" about their memory and the extent to which this knowledge is valid (see Koriat, 1993; Metcalfe & Shimamura, 1994; Nelson & Narens, 1990). Theoretical and empirical treatments of the processing consequences of these abilities have been less common (but see Barnes et al., 1995; Gruneberg et al., 1977; Koriat et al., 1988; Mazzoni & Cornoldi, 1993; Mazzoni, Cornoldi, & Marchitelli, 1990; Metcalfe, 1993; Nelson, 1993; Nelson & Leonesio, 1988; Nelson & Narens, 1990; Reder, 1987, 1988; Reder & Ritter, 1992). Such treatments require a distinction between two separate but related functions, *monitoring* and *control* (Barnes et al., 1995; Koriat et al., 1988; Metcalfe, 1993; Nelson & Narens, 1990, 1994). In applying these concepts to free-report memory performance, 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 (see also Barnes et al., 1995). The control decision is assumed to depend on the monitoring output as well as functional incentives and situational demands.

In Figure 1, these concepts are incorporated into a simple model indicating how monitoring and control processes can be used to regulate memory accuracy and quantity performance under free-report conditions. The model is deliberately schematic: It makes no claims whatsoever about the nature of mem-

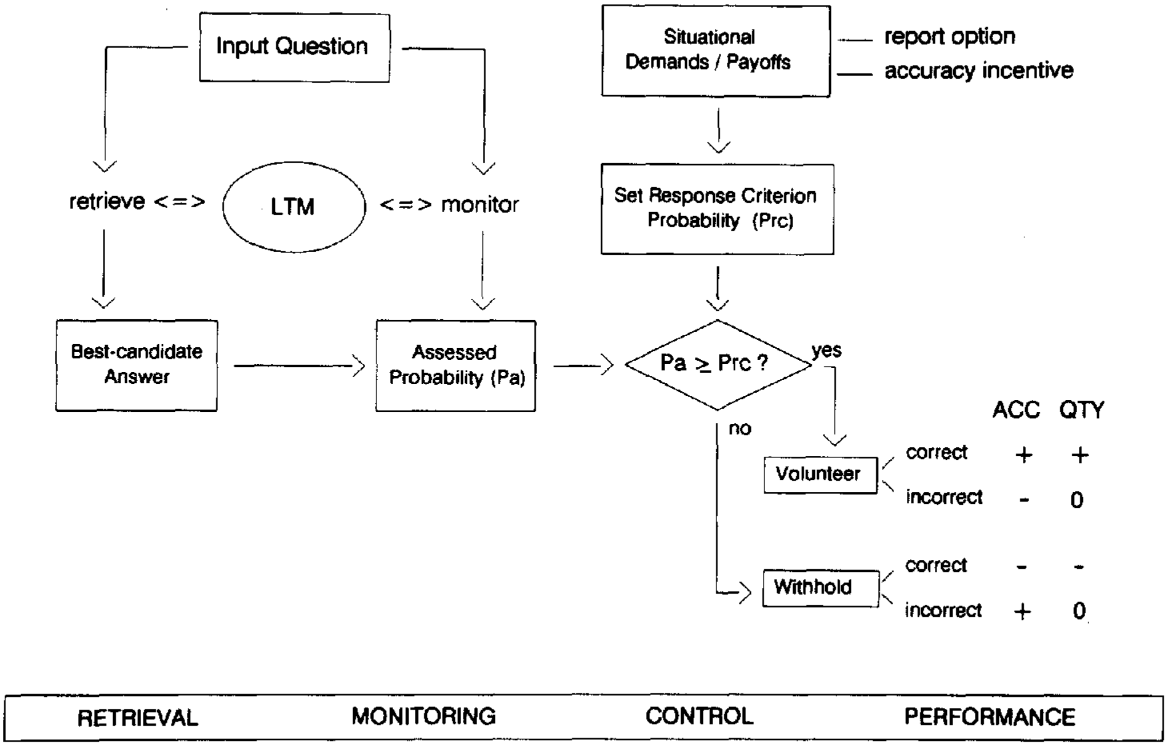

Figure 1. A schematic model of the strategic regulation of memory accuracy and memory quantity performance. Performance effects are signified by plus (increase), minus (decrease), and zero (no effect). LTM = long-term memory; ACC = accuracy; QTY = quantity; P_a = assessed probability; P_{rc} = response criterion probability.

ory retrieval processes, and in fact it may be applied equally well to both recall and recognition. No matter how candidate memory responses are arrived at, our purpose is to investigate the manner in which monitoring and control processes at the reporting stage affect the ultimate memory performance (cf. the distinction between *ecphory* and *conversion* in Tulving, 1983). Thus, we postulate that the combined product of the retrieval (or ecphory) and monitoring processes is a "best-candidate" answer, together with its associated assessed probability, P_a, of being correct. (When no plausible answer is accessible, the best candidate may be any "wild guess," with $P_a = 0$). The control mechanism then compares that assessed probability with a preset response criterion probability, P_{rc}: The answer is volunteered when $P_a \geq P_{rc}$, but withheld otherwise. The P_{rc} threshold is set on the basis of implicit or explicit payoffs, the gains for providing correct answers relative to the costs of giving wrong answers.

Although the assumptions embodied in this model may seem straightforward, the implications for memory performance are not. Within the proposed framework, the contributions of monitoring and control to free-report memory performance can be shown to depend on the following three factors: (a) *monitoring effectiveness*—the extent to which the assessed probabilities successfully differentiate correct from incorrect candidate answers; (b) *control sensitivity*—the extent to which the volunteering or withholding of answers is in fact sensitive to the monitoring output; and (c) *response criterion setting*—the P_{rc} level that is set in accordance with the incentive to be accurate (payoff schedule).

Most previous treatments of the effects of recall criterion, borrowing from signal-detection theory, have focused on the third factor alone (see, e.g., Klatzky & Erdelyi, 1985). Thus, the widely acknowledged prediction is for a quantity–accuracy tradeoff: To the extent that the probability assessments are reasonably diagnostic of the correctness of the candidate answers and the answers are volunteered or withheld on the basis of those assessments, then raising the response criterion will result in fewer volunteered answers, a higher percentage of which are correct (increased accuracy) but a lower number of which are correct (decreased quantity). Because raising the response criterion is assumed to increase accuracy at the expense of quantity, the strategic control of memory performance should require the rememberer to weigh the relative payoffs for accuracy and quantity in reaching an optimal criterion setting.

It has been largely overlooked, however, that both the benefits and the costs of this strategic control, indeed, the very existence of the performance tradeoff, depend critically on the rememberer's monitoring effectiveness. Unlike in the forced-report situation addressed by the signal-detection methodology, in which monitoring effectiveness and memory "retention" are essentially synonymous (see later discussion), under free-report conditions, monitoring effectiveness is quite distinct from both the amount of retention and the criterion level used (Lockhart & Murdock, 1970; and see Experiment 2, later). To illustrate, consider once again an eyewitness who is questioned about a remote event. If the witness' retention is very poor, she might provide very little correct information under directed questioning (i.e., poor forced-report memory-quantity performance). Nevertheless, her monitoring of the correctness of her answers could still be perfect, in which case a free-report interrogation would allow her to volunteer only the few correct details that she trusts, thereby achieving perfect accuracy with no quantity tradeoff (see following simulations). On the other hand, her monitoring could be very poor as well, in which case using the option of free report to screen out answers would not enhance the accuracy of her testimony much or at all (see simulations and Experiment 2 later).

We stress, then, that by taking into account the distinct contributions of memory retention, monitoring, and control, this working model can be shown to yield a rich set of predictions for the effects of each of these factors on both memory-accuracy and memory-quantity performance. In the following section, these predictions will be brought out in simulation analyses based on hypothetical data, following which the model's basic assumptions and predictions will be put to an empirical test.

Simulation Analyses

A series of simulation analyses will now explore the implications of the model just outlined. In all of the simulations, we assume a situation in which a person is tested using a free-production procedure, and for each item a best-candidate answer is associated with one of 11 assessed probability levels (0, .10, . . ., .90, 1.0). The first simulation examines the effects of changes in the response criterion (P_{rc}) on free-report accuracy and quantity performance, assuming a fairly good level of monitoring effectiveness. A second set of simulations then demonstrates how these effects can vary substantially, depending on different qualities of the monitoring output.

Effects of Report Option and Accuracy Incentive

Method

In this analysis, we explore the effects of report option and accuracy incentive on memory performance. For simplicity, we assume that the person's answers are uniformly distributed across the assessed probability categories and that the monitoring output is perfectly calibrated (but see further analyses later). Perfect calibration means that for all answers assigned to a given probability category, the actual proportion of correct answers equals the mean assessed probability, in which case the proportions of correct answers plotted against the mean assessed probabilities all lie along the diagonal line (see Lichtenstein, Fischhoff, & Phillips, 1982). People's monitoring has generally been found to be fairly well calibrated, although a tendency for overconfidence has often been observed (see, e.g., Keren, 1988; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein et al., 1982). Note that under the conditions specified here, 50% of the person's candidate answers are, in fact, correct.

Given this hypothetical data, we simulated the free-report performance that would ensue from the adoption of each of the 11 P_{rc} values: For each value, we first applied the model to the monitoring output to determine which answers to volunteer and which to withhold, and then we calculated the memory accuracy (output-bound proportion correct) and quantity (input-bound proportion correct) measures accordingly.

Results

Figure 2 plots the simulated accuracy and quantity performance as a function of P_{rc}. For $P_{rc} = 0$ (forced report), accuracy and quantity performance are necessarily equivalent, and they simply reflect the amount of overall retention (50%). However, as the response criterion is raised under the option of free report, accuracy increases while quantity decreases (i.e., a quantity–accuracy tradeoff). Note that a quantity–accuracy tradeoff is indicated both in contrasting free- and forced-report performance (report option) and in comparing higher and lower free-report criterion settings (accuracy incentive). Indeed, because forced-report situations require that $P_{rc} = 0$, both of these contrasts may be viewed in terms of criterion level.

In addition to the expected tradeoff pattern, however, the simulated functions also reveal a further, more subtle result: Given the type of well-calibrated monitoring distribution assumed in this simulation, quantity performance decreases as a positively accelerated function of P_{rc}, whereas accuracy increases linearly. Thus, as the response criterion is raised, the rate of the tradeoff increases, and withholding answers to improve accuracy performance becomes relatively costly. This result, which holds across a wide range of distributional and performance assumptions (see next section), may explain an intriguing pattern of results observed in our previous study (Koriat & Goldsmith, 1994). In that study, there was no quantity cost for increased accuracy when comparing free-report and forced-report performance under a moderate incentive (an increase of 19 percentage points in accuracy achieved at an insignificant 3 percentage-point cost in quantity, across recall and recognition), but a substantial quantity cost was evidenced when comparing free-report performance between the high- and moderate-incentive conditions (an additional 15-point increase in accuracy achieved at a 13-point cost in quantity). Thus, in itself, the option of free report allowed a substantial increase in accuracy to be achieved at a negligible cost in quantity but under a stronger accuracy incen-

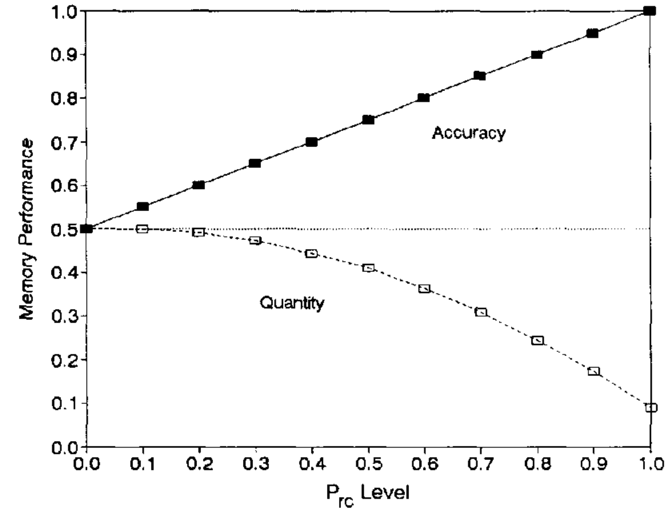

Figure 2. Simulated memory quantity and memory accuracy performance (proportion correct) plotted as a function of P_{rc} level. P_{rc} = response criterion probability.

tive, increasing accuracy even further became disproportionately more costly. As we shall show in the General Discussion, this pattern also emerges when comparing other studies in which the free-report response criterion was manipulated (e.g., the recall-criterion research mentioned earlier).

The Role of Monitoring Effectiveness

We now examine how monitoring effectiveness moderates the effects of report option and accuracy incentive on memory performance. Clearly, the screening of answers on the basis of subjective confidence will enhance accuracy performance only to the extent that the monitoring mechanism can successfully distinguish between correct and incorrect candidate answers. This monitoring ability is commonly referred to as *discrimination accuracy* or *monitoring resolution.* Although various specific measures have been proposed (e.g., Liberman & Tversky, 1993; Murphy, 1973; Nelson, 1984; Schraw, 1995; Yaniv, Yates, & Smith, 1991; Yates, 1982, 1990), for present purposes, resolution may be conceived as reflecting the correlation between the assessed probability assigned to an answer and whether that answer is correct. Resolution is maximized when all correct answers are assigned a high probability and all incorrect answers are assigned a lower probability.[3]

To illustrate the importance of monitoring resolution, let us consider two conditions in which the output of the monitoring mechanism will not be of any use for the control of memory reporting. The first is when there is little or no variance in the probability assessments for different candidate answers, to the extreme that all answers are assigned the same subjective probability. For example, a person who feels incapable of differentiating correct and incorrect answers may therefore assign all answers a probability of .50. Although calibration may be perfect (if the overall proportion correct is indeed .50), resolution is zero. Worse yet, the person's monitoring may be miscalibrated as well, for instance, if the person is convinced that all of his or her answers are correct.

A second type of deficiency occurs when a person does feel that he or she can differentiate between different answers, but there is, in fact, no relationship between the subjective and actual probabilities (see, e.g., Cohen, 1988). As in the first case, resolution is completely lacking, but an important difference is that, unaware of the inadequacy of his or her monitoring, the person may nevertheless control memory reporting in accordance with the inappropriate probability assessments.

These two types of monitoring deficiencies may be seen to represent the worst extremes of two distinct dimensions of monitoring quality. The first dimension, which we shall term *polarization,* concerns the distribution of the probability assessments. The second concerns the *correspondence* between the assessed probabilities and the actual proportions correct. We now examine how these two monitoring dimensions can affect the potential joint levels of memory accuracy and quantity performance.

Method

As in the previous simulation, we assume a free-production procedure in which 50% of the person's best-candidate answers are correct.[4] However, we now manipulate both the polarization and the correspondence of the probability assessments in order to explore the effects of differences in monitoring effectiveness for the ultimate memory performance.

With regard to the polarization dimension, we distinguish three boundary conditions. At one extreme we have the unipolar situation just mentioned, in which the person uses only a single assessed probability (in this case .50) for all answers. At the other extreme, there is a bipolar distribution, in which assessed probabilities are dichotomously confined to the values of 0 and 1.0. These two extremes of the polarization continuum may be seen to conform to two asymptotic standard-form beta distributions (Johnson & Kotz, 1970, p. 37; and see Appendix): The bipolar case is approached as the distributional shape parameters, p and q, both approach zero, and the unipolar case is approached as p and q both approach infinity. Between these two extremes falls the rectangular distribution, in which the probability assessments are uniformly distributed across the domain of assessed probabilities ($p = q = 1$). Note that polarization only pertains to the distribution of the assessed probabilities, without regard to calibration.

As for the correspondence dimension, here we refer to the *slope* of the calibration function for a given distribution of probability assessments. At one extreme, we have the deficiency discussed earlier, in which there is no correspondence between the assessed probabilities and the correctness of the answers, such that the proportion correct (for our hypothetical rememberer) is .50 for all categories. That is, the slope of the calibration function is zero. At the other extreme, we have the case of perfect correspondence, in which the proportions correct are precisely equal to the assessed probability levels, and the calibration plot falls along the diagonal. Note, however, that except in the case of the completely bipolar distribution defined earlier, perfect correspondence still yields less than perfect resolution (see Footnote 3). Intermediate values for this dimension are produced by rotating the calibration function between the slopes of 0 and 1 (see Appendix). We should emphasize that although both calibration and resolution covary along the correspondence continuum, it is the increased resolution that is critical for monitoring effectiveness.

To explore the effects of different qualities of monitoring output, we first computed the P_{rc}–performance functions that would ensue for our hypothetical rememberer, assuming different combinations of the po-

[3] For the unfamiliar reader, it is important to distinguish between calibration and resolution as indexes of monitoring effectiveness (see, e.g., Liberman & Tversky, 1993; Lichtenstein et al., 1982; Yaniv et al., 1991; Yates, 1982, 1990). 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: Although a person 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 resolution (relative correspondence) that is critical for the effective operation of the control mechanism. At the extreme, when assessed probabilities are dichotomously confined to 1.0 (*certainly correct*) and 0 (*certainly incorrect*), perfect calibration entails maximum resolution, resulting in perfect monitoring accuracy at the level of individual items. Here the correlation between assessed probability and the correctness of individual answers is 1.0, but note that the same correlation (and resolution) would also 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.

[4] The analyses to be reported here were also performed while systematically varying the overall performance level. These variations did not change the basic pattern of results.

larization and correspondence boundary conditions just mentioned. We then manipulated both of these dimensions in a more graded manner.

Results

Figure 3 plots the simulated effects of changes in the response criterion on memory accuracy and quantity performance for different combinations of the polarization (d = distribution) and correspondence (s = slope) conditions discussed earlier.

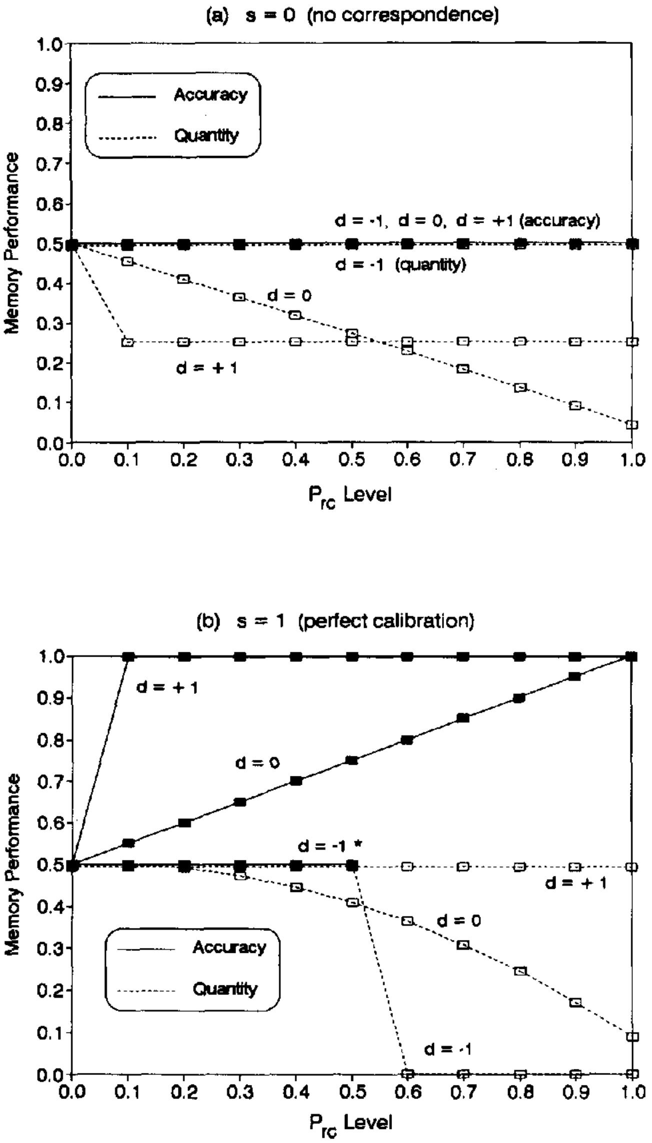

Figure 3. Simulated memory quantity and memory accuracy performance (proportion correct) plotted as a function of P_{rc} level, for different qualities of monitoring output. Panel a plots the performance for three levels of monitoring polarization (unipolar: $d = -1$, uniform: $d = 0$, and bipolar: $d = +1$) when monitoring correspondence is completely lacking ($s = 0$). Panel b plots the performance for the same three levels of polarization when correspondence is perfect ($s = 1$). The asterisk indicates that for $d = -1$, accuracy is undefined when $P_{rc} > .50$ (no answers are volunteered). P_{rc} = response criterion probability.

Consider first the top panel, (a), which depicts the patterns of performance for different levels of polarization when there is no correspondence between assessed probability and proportion correct ($s = 0$). Here volunteered and withheld answers are equally likely to be correct, and therefore exercising the option of free report does not improve accuracy and is only detrimental to quantity performance (for an experimental approximation to this situation, see Experiment 2 later). Thus, in the case of a uniform distribution ($d = 0$), raising the criterion level and withholding more answers steadily reduces quantity performance while leaving accuracy unchanged. When the distribution of assessed probabilities is completely bipolarized ($d = 1$), a fixed drop in quantity performance ensues from withholding the "certainly incorrect" group of items. The unipolar condition ($d = -1$), in order to be miscalibrated, was defined as the assessment of all answers as "certainly correct" ($P_a = 1.0$). Thus, all answers are volunteered, and both accuracy and quantity remain fixed at .50.

Panel b of Figure 3 depicts the expected performance when there is perfect correspondence between assessed probability and proportion correct (perfect calibration, $s = 1$). Compared with Panel a, this condition seems to better approximate the actual monitoring abilities that people generally exhibit (see Experiment 1 later). Even here, however, we see that the correspondence is only meaningful when there is at least some variability in the distribution of assessed probabilities. Thus, the performance function for the unipolar condition ($d = -1$) seems to fit better with Panel a: Either all of the answers are withheld or all are volunteered, but in neither case is accuracy increased beyond .50. Note that memory accuracy cannot even be estimated when no answers are volunteered (indicated on the graph by an asterisk), a problem that does not exist for measuring memory quantity (cf. the "don't know" state in Koriat & Lieblich, 1974, 1977).

With a uniform distribution of assessed probabilities ($d = 0$), on the other hand, we have the same performance pattern that we examined in the earlier simulation (see Figure 2): Raising the response criterion increases accuracy while decreasing quantity (i.e., a quantity-accuracy tradeoff). Note that the drop in quantity performance is less steep than that associated with the corresponding function for $d = 0$ in Panel a. However, as Panel b clearly shows, there is yet a more optimal condition, which is achieved when the perfectly calibrated probability assessments are also completely polarized ($d = 1$), entailing *perfect monitoring resolution.* Here the option of free report allows volunteering only correct and withholding only incorrect answers, and 100% accuracy is achieved across the entire range of free-report criterion levels without any decrease in quantity performance (i.e., no tradeoff).

As indicated earlier, both d and s can be manipulated in a continuous manner between these boundary conditions. It can then be shown that when calibration is held constant, monitoring effectiveness (resolution) steadily improves as d increases from -1 to $+1$. Similarly, holding polarization constant, monitoring effectiveness (calibration and resolution) steadily improves as s increases from 0 to 1 (except for $d = 1$, where there is no resolution in any case). The performance consequences are depicted in Figure 4. The procedures used to manipulate the two dimensions are described more fully in the Appendix.

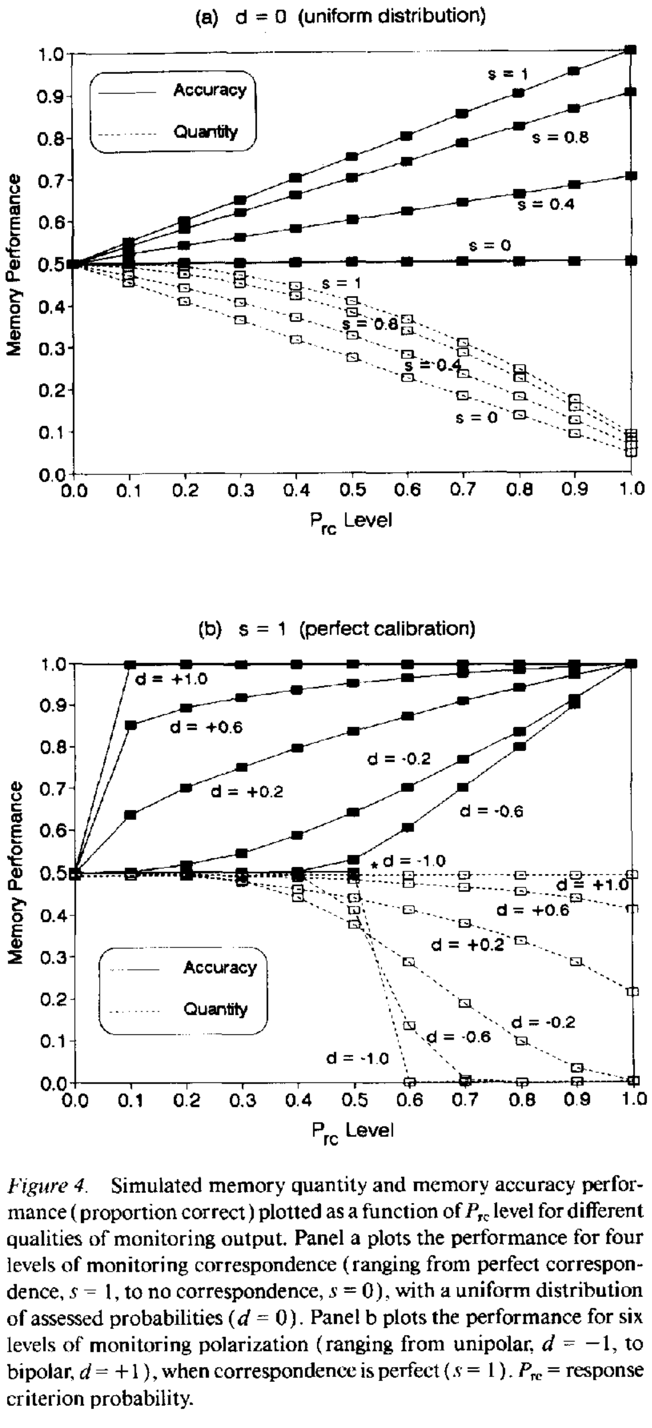

Figure 4. Simulated memory quantity and memory accuracy performance (proportion correct) plotted as a function of P_{rc} level for different qualities of monitoring output. Panel a plots the performance for four levels of monitoring correspondence (ranging from perfect correspondence, $s = 1$, to no correspondence, $s = 0$), with a uniform distribution of assessed probabilities ($d = 0$). Panel b plots the performance for six levels of monitoring polarization (ranging from unipolar, $d = -1$, to bipolar, $d = +1$), when correspondence is perfect ($s = 1$). P_{rc} = response criterion probability.

Implications for Memory Performance

The foregoing simulation analyses bring out several implications regarding the effects of monitoring and control processes on memory performance. In general, report option should yield opposing effects on memory accuracy and memory quantity performance. The magnitudes of these effects, however, are expected to depend on the setting of the free-report criterion in accordance with the operative accuracy incentive: Higher criterion settings should generally yield better accuracy performance, but the corresponding cost in quantity performance may tend to increase in relative terms as the response criterion is raised (see Figure 4).

Furthermore, the effective regulation of memory accuracy should be heavily dependent on monitoring effectiveness. At least some degree of correspondence between subjective judgments and the correctness of the answers (resolution) is required for using the option of free report in order to improve accuracy. In addition, as resolution increases, accuracy can be enhanced at lower costs in quantity. Thus, under conditions of free report, the joint levels of accuracy and quantity performance should depend on both monitoring resolution and the setting of the control criterion.

Of course, it remains to be seen whether people really do control their memory performance in accordance with their monitoring output and whether actual memory performance is in fact sensitive to both monitoring effectiveness and the response criterion. Thus, in the following sections, we report two experiments that exploit the framework depicted in Figure 1 and put some of the basic predictions that emerged in the simulation analyses to an empirical test. Experiment 1 investigates the operation of monitoring and control processes in mediating memory performance. Experiment 2 then examines how differences in monitoring effectiveness may have crucial consequences for the strategic regulation of memory accuracy.

Experiment 1

Experiment 1 used a special procedure that combines both free and forced reporting. A general-knowledge test was administered to participants in either a recall or a recognition format. The participants first 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 or a high accuracy incentive.

This design enables us to trace the links postulated by the model (see Figure 1) between "retrieval" (or "ecphory"; see Tulving, 1983), monitoring, control, and memory performance (accuracy and quantity). First, we tap retrieval (recall or recognition) by treating the forced-report answers provided in Phase 1 as representing the participant's "best-candidate" response for each item. Second, we tap monitoring by treating the confidence ratings as representing the assessed probability (P_a) associated with each best-candidate answer. This allows monitoring effectiveness to be evaluated. Third, we tap the control mechanism by examining which answers are volunteered or withheld on Phase 2. This allows us to determine the sensitivity of the control policy to the monitoring output and to derive an estimate of the response criterion (P_{rc}) set by each participant. In addition, comparison of the estimated criterion levels for the two incentive conditions allows examination of the predicted effects of accuracy incentive on the participants' control policy. Finally, the design allows us to evaluate the contribution of monitoring and control processes to the ultimate free-report memory accuracy and memory quantity performance in both recall and recognition test formats. Test format was included because this factor has traditionally been confused with report option and because test-format comparisons can be used to illustrate possible differences in the way that monitoring and control processes are used in the strategic regulation of free-report memory performance.

Method

Participants

University of Haifa undergraduates ($N = 71$), 18 men and 53 women, participated in the experiment for course credit and the chance to win up to New Israeli Shekel (NIS) 60 (approximately $30). They were randomly assigned to each of four Test Format × Incentive conditions, with 17 to 19 participants in each group.

Stimulus Materials and Procedure

Two versions of a 60-item general-knowledge test (in Hebrew) developed by Koriat and Goldsmith (1994, Experiment 1) were used: a recall version and a 5-alternative multiple-choice recognition version. The questions for the two tests were identical, but in the recall version a blank line was provided next to each question for recording the response, whereas in the recognition version the correct answer plus four foils were listed for selection. (The foils were designed to be as plausible as possible.) The questions were formulated such that the correct answer was always a single word or a proper name (cf. Brown & McNeill, 1966; Nelson & Narens, 1990). Examples of the questions used are, What was the name of the first emperor of Rome? What is the chemical process responsible for the formation of glucose in the plant cell?

The experiment included two phases:

Phase 1. Participants took either a recall or a recognition version of the test under forced-report instructions: They were required to answer all the questions, even if they had to guess, and they were also required to assess the likelihood that their answer was correct, using a 0–100% scale for recall and a 20%–100% scale for recognition. Participants were urged to use the entire range of estimates. No monetary incentive was offered for performance on this phase.

Phase 2. After completing Phase 1, participants were given the same test again but under free-report instructions: Here they were told that they could choose whether to answer any given question and that they would not be penalized (but neither would they receive any bonus) for omitted items. Accurate responding was induced by one of two payoff schedules (accuracy incentives): In the moderate-incentive condition, participants were paid NIS 1 (approximately 50¢) for each correct answer and penalized the same amount for each incorrect answer, for a random sample of 15 items. In the high-incentive condition, participants were paid NIS 1 for each correct answer but penalized NIS 10 for each incorrect answer, and payment was based on all volunteered answers. Participants were assured that although they might not break even, they would not have to pay any losses.

Results

Memory Accuracy and Memory Quantity Performance

We begin by examining the performance effects of report option and accuracy incentive on accuracy and quantity performance. We then turn to the results regarding the monitoring and control mechanisms postulated to mediate these effects. Table 1 presents the mean memory quantity and accuracy scores for forced report (Phase 1) and free report (Phase 2), by accuracy incentive (Phase 2 only) and test format.

The effects of report option were evaluated by confining the analyses to the moderate-incentive group: First, a Report Option × Test Format analysis of variance (ANOVA) for the accuracy scores yielded significant effects for report option, $F(1, 32) = 206.82$, $p < .0001$; test format, $F(1, 32) = 13.06$, $p < .001$; and the interaction, $F(1, 32) = 6.44$, $p < .05$. Giving participants the option of free report allowed them to improve their accuracy relative to forced report. This improvement was more pronounced for recall than for recognition, but it was significant in both cases, $t(16) = 9.74$, $p < .0001$, and $t(16) = 11.96$, $p < .0001$, respectively.[5] Second, the same ANOVA performed on the quantity scores also yielded significant effects for report option, $F(1, 32) = 66.74$, $p < .0001$; test format, $F(1, 32) = 12.16$, $p < .005$; and the interaction, $F(1, 32) = 13.08$, $p < .001$. Examination of the means reveals a quantity–accuracy tradeoff: The improved accuracy of free report was accompanied by reduced quantity performance relative to forced report. Unlike the accuracy improvement, this reduction was more marked for recognition than for recall, but it too was significant in both cases, $t(16) = 4.30$, $p < .0005$, and $t(16) = 6.95$, $p < .0001$, respectively. Note, then, that the option of free report allowed recall participants to achieve a greater improvement in accuracy performance at a lesser cost in quantity performance than recognition participants, perhaps implying better monitoring effectiveness. This point will be addressed again later.

Turning next to the incentive manipulation, a two-way ANOVA, Incentive × Test Format, on the free-report (Phase 2) accuracy scores yielded significant effects for incentive, $F(1, 67) = 14.77$, $p < .0005$, and test format, $F(1, 67) = 5.28$, $p < .05$, with no interaction. As predicted, the stronger accuracy incentive induced better accuracy performance than the more moderate incentive, and as mentioned earlier (see Footnote 5), recognition accuracy was found to be somewhat better than recall.

On the other hand, the same ANOVA on the quantity measure did not yield a significant incentive effect ($F < 1$), implying that participants were able to increase their accuracy with no cost in quantity. However, participants in the high-incentive condition yielded a higher quantity score (57.3%) than did moderate-incentive participants (52.5%), even on Phase 1. When this initial difference was partialled out in an analysis of covariance, the incentive effect on accuracy remained significant, $F(1, 66) = 12.17$, $p < .001$ (adjusted means: 88.7% for high incentive vs. 81.4% for moderate), and the incentive effect on quantity was now also significant, $F(1, 66) = 13.57$, $p < .0005$ (adjusted means: 39.1% for high incentive vs. 44.6% for moderate). Thus, a quantity–accuracy tradeoff was obtained for the incentive manipulation after all.

[5] We should note that when only the standard memory measures, free recall and forced recognition, are compared, the "recall–recognition paradox" (Koriat & Goldsmith, 1994) discussed earlier was also replicated here (see Table 1): Quantity performance is superior for forced recognition than for free recall, but accuracy performance is superior for free recall than for forced recognition. However, when the effects of test format and report option are unconfounded, accuracy performance is seen to vary with report option, not test format. As discussed earlier, in the forced-report condition, the accuracy and quantity measures are equivalent by definition. In the free-report condition, on the other hand, the two measures are operationally independent, and here recognition accuracy was, if anything, slightly better than recall accuracy.

Table 1
Mean Quantity and Accuracy Memory Scores (Percent Correct) by Accuracy Incentive for Each Test Format in Experiment 1

	Phase 2 (free report)								Phase 1 (forced report)	
	Quantity (% correct)				Accuracy (% correct)					
	Moderate incentive		High incentive		Moderate incentive		High incentive			
Test format	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Recall	38.0	2.8	38.5	3.3	76.4	3.2	88.1	2.1	46.7	2.7
Recognition	47.3	3.0	43.6	2.8	84.4	2.5	91.1	1.6	63.5	1.7

Monitoring Analyses: Subjective Confidence and the Likelihood of Being Correct

We now turn to the metamemory processes assumed to mediate the effects of report option and accuracy incentive just reported, beginning with the monitoring process. In terms of the model (see Figure 1), the confidence ratings elicited in Phase 1 are assumed to reflect the subjective probability (P_a) that the corresponding best-candidate answer is correct. Because these assessments will be used to predict volunteering behavior in Phase 2, it is important to note that participants were quite consistent in the answers they provided to the questions in the two phases: Only 1.1% of the answers differed, and these answers were eliminated from the analyses that follow.

We first examine the correspondence between the assessed and actual probabilities of being correct. Because the incentive manipulation was introduced only in Phase 2, the data were collapsed across the two incentive conditions. Both absolute correspondence (calibration) and relative correspondence (resolution) were evaluated. The calibration data are presented in Figure 5 for the recall and recognition conditions, plotted according to the procedure commonly used in calibration studies (see Lichtenstein et al., 1982): The probability judgments are grouped into 10 levels for recognition (.20, .21–.30, .31–.40, . . . , .91–.99, 1.0) and 12 levels for recall (0.0, .01–.10, .11–.20, . . . , .91–.99, 1.0). The proportion correct is plotted against the mean assessed probability, computed across participants; perfect calibration is indicated by the diagonal line.

The plots show a strong positive relationship between mean assessed probability and actual likelihood of being correct for both recall and recognition. The general pattern of deviation from the diagonal is consistent with previous calibration studies (see Erev, Wallsten, & Budescu, 1994), but participants' assessments were fairly well calibrated overall. The proportion correct averaged .47 for recall and .64 for recognition, whereas the respective values for the assessed probabilities were .50 and .67, respectively. Calibration scores for each participant, computed as the weighted mean of the absolute differences between the mean assessed probability and actual proportion correct for each category (Oskamp, 1962), averaged .13 for both the recall and the recognition groups.

As noted earlier, however, it is monitoring resolution that is crucial for successful memory accuracy performance (see Footnote 3). Two measures of resolution were calculated, the Kruskal–Goodman gamma correlation commonly used in metamemory research (see Nelson, 1984, 1986; but see Schraw, 1995; Swets, 1986, for reservations) and the adjusted normalized discrimination index (ANDI) recommended more recently by Yaniv et al. (1991). Both measures indicated good levels of monitoring effectiveness: First, the gamma correlation between confidence and the correctness of each answer averaged .87 for the recall participants (range = .65–1.0), significantly different from zero, $t(35) = 65.3$, $p < .0001$, and .68 for the recognition participants (range = .44–.93), also significantly different from zero, $t(34) = 33.4$, $p < .0001$. According to this index, the recall participants were somewhat more effective in monitoring the correctness of their answers than were the recognition participants, $t(69) = 7.96$, $p < .0001$.

Second, we calculated the ANDI measure, which has a straightforward interpretation as the proportion of variance in the correctness of the answers that is accounted for by the participant's probability judgments. ANDI is not biased by differences in the overall level of memory performance (neither is gamma; see Nelson, 1984) or in the number of probability categories used. This index averaged .61 for recall (range = .29–.92) and .30 for recognition (range = .08–.59), $t(69) = 9.37$, $p < .0001$, for the difference.

According to both analyses, then, participants were quite effective in discriminating correct from incorrect answers, but the recall participants were more effective than the recognition participants. This latter difference appears to derive from the polarization dimension discussed earlier. As indicated in Figure 5, although the recall and recognition participants were equally calibrated, the recall monitoring was more polarized than the recognition monitoring: Whereas both the recall and the recognition participants used the highest (1.0) probability category about equally often ($n = 618$ and $n = 678$, respectively), the recall participants were far more likely to use their lowest probability category ($n = 635$) than were the recognition participants ($n = 273$). Cast in percentage terms, 58% of the recall assessments fell into the two most extreme categories, compared with 46% for recognition. Hence, given equal calibration, the more polarized recall monitoring should and did yield bet-

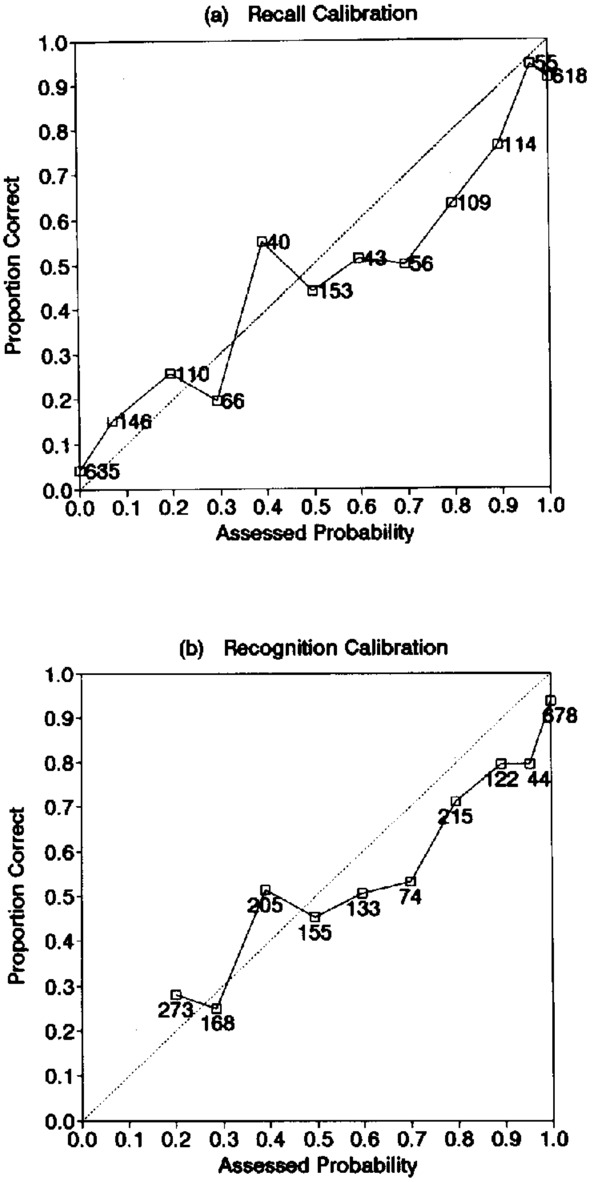

Figure 5. Calibration curves for the recall and recognition participants in Experiment 1. The frequency of judgments in each category appears beside each data point.

ter monitoring effectiveness.[6] We will return to consider the implications of this difference later.

Control Analyses: Subjective Confidence and the Decision to Respond

We now turn to the results pertaining to the operation of the control mechanism, which is assumed to determine whether the best-candidate answer will be volunteered or withheld. According to the model, this decision is based on (a) the output of the monitoring mechanism and (b) the incentive for accuracy.

The contribution of monitoring. Overall, the results suggest that the decision to volunteer an answer does indeed depend heavily on one's confidence in the correctness of that answer. Across all participants and conditions, items that were volunteered in Phase 2 had a mean assessed probability of .89 in Phase 1, compared with .29 for items that were withheld. In addition, the relationship between confidence and volunteering was calculated using both gamma and ANDI, and by both measures it was found to be exceedingly strong: Gamma averaged .97 for recall participants (range = .89–1.0) and .93 for recognition (range = .68–1.0). Similarly, ANDI (the proportion of explained variance) averaged .79 for recall (range = .52–1.0) and .67 for recognition (range = .39–.93). The difference between recall and recognition was relatively small but significant for both measures, $t(69) = 3.23$, $p < .0005$, and $t(69) = 4.17$, $p < .0001$, respectively.

Note that the confidence-volunteering correlations reported here were higher than the confidence–correctness correlations reported earlier. For example, whereas the gamma correlations between confidence and the decision to volunteer an answer averaged .97 and .93 for recall and recognition, respectively, the corresponding gammas for the confidence–correctness relationship averaged .87 and .68 (the former correlations were significantly higher than the latter, both for recall, $t[35] = 6.86$, $p < .0001$, and for recognition, $t[34] = 7.47$, $p < .0001$). Thus, in exercising the option of free report, participants seem to have put more faith in their subjective confidence than was actually warranted, presumably because they had no better predictor (see Experiment 2).

The effect of incentive. We now examine the effect of accuracy incentive on volunteering behavior and show how this effect can be captured in terms of response criterion. As predicted, the high-incentive participants volunteered fewer answers (26.9) than did the moderate-incentive participants (30.9), $t(69) = 2.0$, $p < .05$. According to the model, this difference reflects the setting of a higher response criterion (P_{rc}), and in fact, mean confidence for volunteered items was .93 for the high-incentive participants compared with .84 for the moderate-incentive participants, $t(69) = 4.70$, $p < .0001$.

In addition, a computational procedure was used to estimate each participant's criterion level. Considering each probability rating actually used by the participant as a candidate P_{rc}, we

[6] It should be pointed out that there is an inherent constraint on recognition monitoring resolution due to the baseline probability of correctly guessing an answer. For instance, in a five-alternative multiple-choice test, one can only try to distinguish between those answers that are "certainly correct" and those that have a 20% chance of being correct. Thus, it is generally not possible to completely differentiate correct and incorrect answers. In a somewhat different context, Schwartz and Metcalfe (1994) noted that differences in baseline probabilities can complicate the comparison of monitoring accuracy for feeling-of-knowing judgments (following unsuccessful recall) across different criterion tests. In the present context, however, multiple-choice baseline probabilities are best viewed as an intrinsic aspect of recognition monitoring, tending to increase the quantity cost of selective reporting, but significantly, not necessarily constraining potential free-recognition accuracy (see Figure 7 later). We should also note that in line with our results, Dunlosky and Nelson (in press) found judgments of learning to be both more polarized and more accurate when elicited by cued-recall probes than by recognition probes, even though the criterion test was the same (recognition) in both cases.

Table 2
Mean Estimated P_{rc} Values for Each of the Four Experimental Conditions in Experiment 1

	Test format							
	Recall				Recognition			
	Moderate incentive		High incentive		Moderate incentive		High incentive	
Variable	$P_a \geq P_{rc}$	$P_a < P_{rc}$	$P_a \geq P_{rc}$	$P_a < P_{rc}$	$P_a \geq P_{rc}$	$P_a < P_{rc}$	$P_a \geq P_{rc}$	$P_a < P_{rc}$
Mean no. responses volunteered	27.1	2.4	24.5	1.1	28.1	4.1	25.6	2.7
Mean no. responses withheld	1.8	28.0	1.9	32.3	2.9	23.3	1.9	29.3
P_{rc}	.50		.80		.71		.88	
Fit ratio	.93		.95		.88		.92	

Note. P_a = assessed probability; P_{rc} = response criterion probability; Fit ratio = the proportion of responses conforming to the decision rule: If $P_a \geq P_{rc}$ then volunteer; else withhold.

defined *hits* as volunteered answers for which $P_a \geq P_{rc}$ and *correct rejections* as withheld answers for which $P_a < P_{rc}$. *False alarms* and *misses* were also defined accordingly. The chosen P_{rc} estimate for each participant was that value that maximized the proportion of hits and correct rejections combined (the *fit ratio*). Because any assessed probability between the chosen P_{rc} and the next lowest assessed probability actually used by the participant would yield an equivalent fit ratio, the derived P_{rc} estimate is actually an upper bound on the estimate for the true criterion level.

Table 2 presents the mean estimated P_{rc}, the mean number of hits, correct rejections, false alarms, and misses, and the fit ratio, for each of the four experimental conditions. Overall, the hits and correct rejections combined accounted for 92% of the participants' responses. Thus, the assumption of a control mechanism operating as a threshold on the monitoring output appears to yield a good approximation of participants' actual volunteering behavior. Moreover, comparing the estimated P_{rc} values for the two incentive conditions, the mean P_{rc} for participants in the high-incentive condition (.84) was significantly higher than for participants in the moderate-incentive condition (.61), $t(69) = 4.81$, $p < .0001$. Thus, as predicted, the high-incentive participants used a stricter control policy than did the moderate-incentive participants. Note that the normative criterion settings for the high- and moderate-incentive conditions are .91 and .50 respectively.[7]

Possible effects of test format. Analysis of the estimated P_{rc} values also suggests that the recognition participants were more conservative than the recall participants in their control policies. P_{rc} averaged 65.9 and 79.6 for recall and recognition, respectively, $t(69) = 2.58$, $p < .05$. The performance consequences of this difference, however, are complicated by the fact that mean confidence for the recognition group (.67) was higher than for recall (.50). Therefore, even though the recognition participants adopted a more conservative criterion for responding, they nevertheless volunteered as many answers (30.2) as did the recall participants (27.5), $t(69) = 1.37$, *ns*.

Thus, the comparison of recall and recognition performance in this experiment reveals a complex interplay between monitoring, control, and the overall correctness of candidate answers. It was shown earlier that recall participants were more effective in monitoring the correctness of their answers. We now see that recognition participants were more conservative in their control policy,[8] although the increased correctness of their candidate answers allowed them to volunteer as many answers as did the recall participants. The net result appears to be an advantage in recognition quantity performance, achieved at no disadvantage in accuracy compared with recall (see also Koriat & Goldsmith, 1994; and see Figure 7, later).

Discussion

Experiment 1 examined the manner in which monitoring and control processes contribute to memory accuracy and quantity performance. The results were generally consistent with the model: First, participants were successful but not perfect in monitoring the correctness of their answers. Second, the tendency to report an answer under free-report conditions was very strongly correlated with subjective confidence in the correctness of the answer. Third, this tendency was also sensitive to

[7] We also conducted several analyses comparing participants' control policies and the ensuing performance to two other control policies: (a) a *normative* control policy, defined as that P_{rc} setting that ensures that all (and only) answers with a nonnegative expected utility (assuming perfect calibration) are volunteered, and (b) an *optimal* control policy, defined as that P_{rc} setting which, when applied to each participant's actual monitoring output, yields the maximum possible performance payoff for that participant. The results will not be reported here, except to note that participants were found to be relatively effective in choosing a control policy that would maximize their performance in accordance with the specified payoff schedule.

[8] Note that a common belief (expressed, for instance, by an anonymous commentator on an earlier draft; see also, Murdock, 1974, p. 65, Figure 3.8) is that the response criterion underlying recall testing is inherently more strict than that underlying recognition responding. Our results, in contrast, suggest that when test format and report option are unconfounded, the opposite might actually be the case. Clearly, more work is needed on this matter.

accuracy incentive: High-incentive participants adopted a stricter criterion than did moderate-incentive participants. Fourth, the exercise of strategic control resulted in improved accuracy performance but also in reduced quantity performance. Finally, consistent with the simulation analyses, the quantity cost of the improved accuracy increased in relative terms when a higher criterion was used: 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).

In sum, the results of this experiment reinforce the initial simulations in demonstrating the contributions of monitoring and control processes to memory performance in an actual memory situation. In Experiment 2, we examine these contributions further under different levels of monitoring effectiveness.

Experiment 2

In Experiment 1, people were found to be quite successful in regulating their free-report memory performance. According to the proposed framework, however, this success should be strongly contingent on monitoring effectiveness. As demonstrated in the simulation analyses, without at least a fair degree of monitoring resolution, the control of memory reporting might not enhance memory performance much or at all.

Several reports in the literature indicate situations in which memory monitoring may be rather poor: First, monitoring may be impaired for certain stimulus materials. Cohen (1988), for example, found that although people were quite accurate in monitoring the recallability of studied words, their judgments of the recallability of self-performed tasks had no predictive validity whatsoever (for a somewhat different example, see Metcalfe & Wiebe, 1987). Also, Fischhoff, Slovic, and Lichtenstein (1977) showed that certain so-called "deceptive" general-knowledge questions tend to produce an illusion of knowing, engendering an undue confidence in one's incorrect answers. In fact, Koriat (1995a) recently found that when people failed to recall the answer to such deceptive items, their feeling-of-knowing judgments were either uncorrelated or even negatively correlated with subsequent recognition memory performance. Second, monitoring may be impaired by various interventions. As a prominent example, exposure to postevent misinformation has also been shown to engender a dissociation between confidence and the validity of people's answers, causing witnesses, for instance, to report with high confidence that they saw three perpetrators of a crime instead of the actual two (Weingardt, Leonesio, & Loftus, 1994; see also Chandler, 1994). Finally, there is evidence indicating that monitoring abilities may be relatively impaired in certain special populations, for example, young children (e.g., Pressley, Levin, Ghatala, & Ahmad, 1987); Korsakoff patients (e.g., Shimamura & Squire, 1986); and patients with frontal lobe lesions (e.g., Janowsky, Shimamura, & Squire, 1989).

According to the model, such monitoring deficiencies should have crucial consequences for the strategic regulation of memory performance (cf. Bjork, 1994). It is our assumption that people have no direct access to the correctness of their answers (Koriat, 1993) and hence have no better choice than to trust their own subjective confidence in controlling their memory reporting. Therefore, when monitoring is poor, selective reporting should yield relatively small accuracy gains accompanied by relatively large quantity costs (see Figure 4). At the extreme, the withholding of answers on the basis of invalid subjective probabilities could be entirely detrimental: It could fail to improve memory accuracy performance and only reduce memory quantity performance (see Figure 3, Panel a).

Experiment 2 was designed to examine these ideas by comparing the effects of free-report control processes under two different levels of monitoring effectiveness. To manipulate monitoring, two samples of general-knowledge items were used: The first consists of "deceptive" items of the sort used by Fischhoff et al. (1977; see also May, 1986), which, as mentioned earlier, tend to elicit a relatively large proportion of incorrect answers that may nonetheless be held in high confidence. For instance, when asked to name the capital of Australia, many people confidently report Sydney rather than Canberra; or, when asked to name the composer of the "Unfinished Symphony," they may confidently report Mozart or Beethoven rather than Schubert. Whatever the reasons for such illusions of knowing (see Fischhoff et al., 1977; Koriat, 1995b), we expect participants' confidence judgments for these items to be generally undiagnostic of the correctness of their answers. The second sample consists of "standard" or typical items that are expected to yield a relatively good level of monitoring effectiveness, like that observed in Experiment 1. In addition, a third set of "difficult" items was also included, comprising items that seldom bring to mind any answers at all. As will be explained later, these items were included so that when combined with the standard items, they would allow comparisons in which forced-report performance for the standard and deceptive conditions is equated.

In this experiment, then, recall participants were given a 90-item general-knowledge test in which the standard, deceptive, and difficult items were randomly intermixed. The procedure was basically the same as in Experiment 1, including both forced-report and free-report phases. Assuming that monitoring resolution is relatively good for the standard items but poor for the deceptive items, we expect that (a) participants will nevertheless base their control decisions on their monitoring output in a similar manner for both sets of items; (b) whereas for the standard items, the option of free report will result in a substantial accuracy increase, for the deceptive items the accuracy increase will be negligible; and (c) a much greater quantity–accuracy tradeoff will be evidenced for the deceptive items than for the standard items. These predictions are expected to hold even when the standard and deceptive conditions are equated in terms of forced-report memory performance.

In addition, unlike in Experiment 1, in which the free-report phase always followed the forced-report phase, in Experiment 2 the order of the two phases was counterbalanced across participants. This allowed us to examine whether the results of Experiment 1 hold across both phase orders.

Method

Participants

University of Haifa undergraduates ($N = 30$), 7 men and 23 women, participated in the experiment. They were paid NIS 25 (approximately $8) for participation and were given the chance to win up to an additional NIS 90 (approximately $30). They were randomly assigned to each of the two phase orders. The experiment was administered in group sessions lasting about 1 hr.

Stimulus Materials

A 90-item general-knowledge test was compiled in a recall format. As in Experiment 1, the correct answer for all items was either a single word or a proper name. In designing the test, three subsets of 30 items were selected on the basis of norms collected by Koriat (1995a). The first, deceptive subset (cf. Fischhoff et al., 1977) included items that had been found to evoke unwarranted feeling-of-knowing (FOK) judgments following unsuccessful recall (e.g., "Who composed the 'Unfinished Symphony?' "). The FOK judgments for these items were generally uncorrelated or even negatively correlated with subsequent recognition memory performance. A second, standard subset was selected that was approximately equated with the set of deceptive items in recall accessibility (i.e., the likelihood that some answer would be volunteered; see Koriat, 1995a). Unlike the deceptive items, however, these items had been found to yield generally valid FOK judgments. The third, difficult subset consisted of items that had been found to elicit few candidate answers at all. This set was included because mean forced-report quantity performance is expected to be higher for the standard items than for the deceptive items, and therefore these difficult items will be combined with the standard items in special analyses that equate forced-report performance for the standard and deceptive subsets. Items from each of the three subsets were randomly intermixed among the 90 test items.

Procedure

As in Experiment 1, participants were given the same test twice, in both forced-report and free-report phases. The order of the two phases was counterbalanced across participants:

Forced-report phase. In this phase, participants were required to answer all the questions, even if they had to guess, and to indicate their confidence in the correctness of the provided answer on a 0–100% scale. No monetary incentive was offered for performance on this phase.

Free-report phase. In this phase, participants were told that they could choose whether to answer any given question and that they would not be penalized (but neither would they receive any bonus) for omitted items. Accurate responding was induced by a moderate-incentive payoff schedule similar to that of Experiment 1: Participants were paid NIS 1 for each correct answer and penalized the same amount for each incorrect answer. Participants were assured that although they might not break even, they would not have to pay any losses, nor would they forfeit any amount of their fixed payment for participation.

Results

Effects of Phase Order

To determine whether the basic pattern revealed in Experiment 1 was sensitive to phase order, we first examined the effects of phase order on both memory and metamemory performance for the standard items only. In general, phase order had little or no effect, and the same pattern of results obtained in Experiment 1 was replicated here for each phase order considered separately. In addition, with regard to the comparisons between the two monitoring conditions, standard and deceptive, that are of primary interest in this experiment, in only one of the analyses to be reported later was there a significant interaction involving phase order. Thus, although phase order was included as a factor in all of the following ANOVAs comparing the two monitoring conditions, results pertaining to phase order will be reported only in that one case in which they are relevant. Also, as in Experiment 1, answers that differed between the forced-report and free-report phases (less than 0.8% of all answers) were eliminated from the analyses.

Analysis of Monitoring Effectiveness

We first examine whether monitoring effectiveness indeed differed for the standard and deceptive items. Calibration curves based on the forced-report phase were computed as in Experiment 1 and are plotted separately in Figure 6 for the standard and deceptive subtests.

The difference between the two monitoring conditions is immediately apparent: Whereas for the standard items, confidence in an answer was generally diagnostic of its correctness, for the deceptive items, the relationship between assessed probability and actual proportion correct is negligible. First, in terms of calibration, participants were much more overconfident on the deceptive subtest than on the standard subtest: For the deceptive subtest, confidence averaged .32 when the actual proportion correct was .12, whereas the respective values for the standard subtest were .31 and .28. Also, the participants' calibration scores, computed as the weighted mean of the absolute differences between the mean assessed probability and actual proportion correct for each category, averaged .13 for the standard subtest (same as Experiment 1) but .26 for the deceptive subtest. Second, and more important, the difference between the two conditions in terms of monitoring resolution is no less dramatic: Gamma averaged .90 for the standard subtest versus .26 for the deceptive subtest, and ANDI averaged .64 for the standard subtest versus .03 for the deceptive subtest. In sum, the choice of standard and deceptive items appears to have provided a very successful manipulation of monitoring effectiveness: Whereas the standard items yielded a relatively good level of monitoring, closely resembling that of Experiment 1, monitoring for the deceptive items was by all accounts very poor.

Analysis of Control

Notwithstanding the success of the monitoring manipulation, we expected that participants would necessarily base their selective reporting on their monitoring output, regardless of its validity. Indeed, for both the standard and the deceptive subtests, there was a very strong relationship between confidence and volunteering: For the standard subtest, volunteered items had a mean assessed probability of .83, compared with .10 for items that were withheld, and the respective means for the deceptive subtest were .67 versus .15.

The confidence-volunteering relationship for each subtest

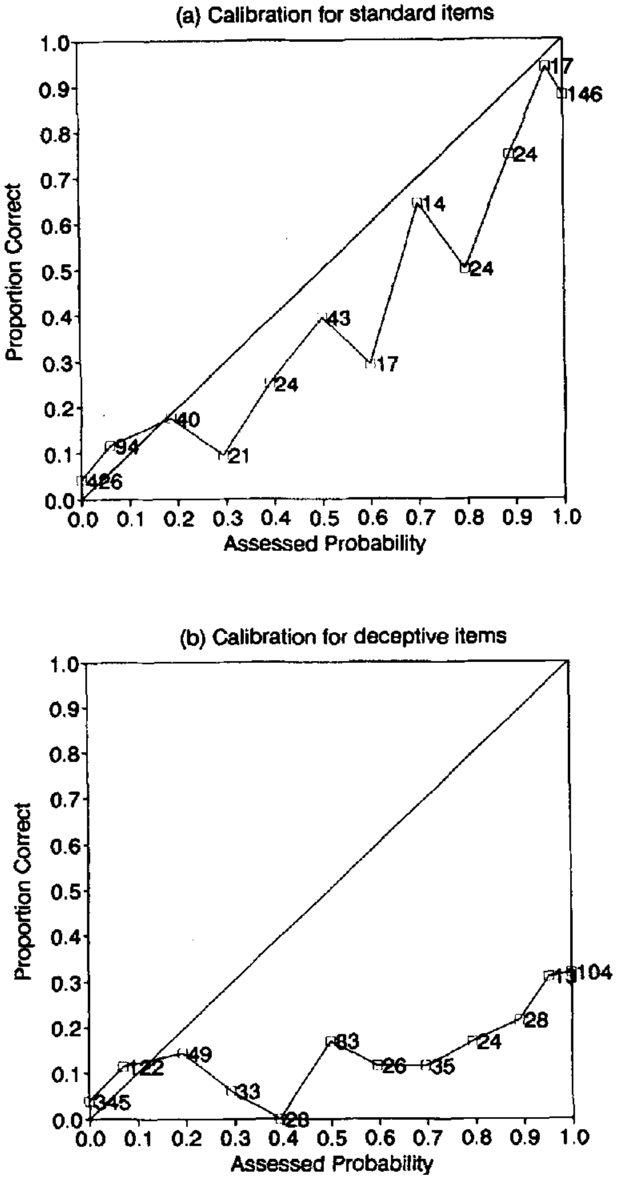

Figure 6. Calibration curves for the standard and deceptive monitoring conditions in Experiment 2. The frequency of judgments in each category appears beside each data point.

was evaluated further using both gamma and ANDI. Gamma averaged .95 for the standard subtest and .88 for the deceptive subtest. The respective values for ANDI were .81 and .58. Thus, volunteering decisions on both subtests were extremely sensitive to the assessed probabilities but somewhat more so on the standard subtest than on the deceptive subtest: The difference was significant both for gamma, $F(1, 28) = 7.26, p < .05$, and for ANDI, $F(1, 28) = 20.99, p < .0001$. The somewhat stronger relationship found for the standard subtest was unexpected. Perhaps confidence for the standard items was more stable across the forced-report and free-report phases than it was for the deceptive items, although at present this is only speculation.

Did participants use a similar control policy for the standard and deceptive items? Response criterion (P_{rc}) estimates, derived as in Experiment 1, averaged .68 and .63 for the standard and deceptive subtests, respectively, indicating that participants used equivalent response criteria for both types of items ($F < 1$). The effect of phase order, however, was significant, $F(1, 28) = 4.53, p < .05$, as was the interaction, $F(1, 28) = 5.58, p < .05$: Across both subtests, participants in the forced-free condition used a higher criterion (.75) than did free-forced participants (.56), although the difference was significant only for the deceptive subtest (.79 vs. .47, respectively), not for the standard subtest (.71 vs. .65, respectively). Thus, although there was no systematic difference in the criterion levels adopted for the standard and deceptive items, giving participants the option of free report after the forced-report phase may have elicited a somewhat greater tendency for selective reporting than when free reporting preceded forced.

Performance Consequences

Given the similar control policies but the different monitoring effectiveness exhibited for the standard and deceptive subtests, what are the implications for actual memory performance? The memory performance measures for the two conditions are presented in Table 3.

The results for the standard subtest disclose a pattern very similar to that obtained in Experiment 1: The option of free report allowed participants to increase their accuracy performance substantially compared with their performance on forced report, $F(1, 28) = 106.76, p < .0001$, and this improvement was achieved at a small cost in quantity performance, $F(1, 28) = 25.22, p < .0001$. Indeed, the free-report accuracy obtained on the standard subtest (75.0%) was virtually identical to that demonstrated by the moderate-incentive recall participants of Experiment 1 (76.4%). In contrast, for the deceptive, poor-monitoring items, participants were able to achieve on the average no better than 21% accuracy when given the option of free report! (Recall that according to the payoff scheme, participants needed to achieve at least 50% accuracy just to break even.) This level of free-report accuracy was only slightly better than forced-report accuracy, $F(1, 28) = 11.08, p < .005$.

Furthermore, as predicted, the quantity–accuracy tradeoff was far more severe under conditions of poor monitoring: Whereas the accuracy improvement for the standard subtest (47 percentage points) was much greater than that achieved on the deceptive subtest (9 percentage points), $F(1, 28) = 48.44$, $p < .0001$, the quantity cost was equivalent (6 and 4 percentage points, respectively), $F(1, 28) = 1.49$, *ns.* (Note that the quan-

Table 3
Mean Quantity and Accuracy Memory Scores (Percent Correct) by Monitoring Conditions in Experiment 2

Monitoring condition	Free report (% correct)				Forced report (% correct)	
	Quantity		Accuracy			
	M	*SE*	*M*	*SE*	*M*	*SE*
Standard	22.3	3.7	75.0	4.7	27.9	3.8
Deceptive	7.6	2.1	21.0	3.9	11.8	2.2

tity cost exhibited for the deceptive subtest was also significant, $F[1, 28] = 26.67, p < .0001$.)

As anticipated, however, there is one difference between the standard and deceptive subtests that complicates interpretation of the results. Forced-report performance for the standard subtest was substantially better than for the deceptive subtest, $F(1, 28) = 62.87, p < .0001$, which means that in addition to the intended difference in monitoring effectiveness, the standard subtest was also easier. Conceivably, this difference could be responsible for some of the effects just reported. To evaluate this possibility, then, we used items from the *difficult* subset in conjunction with items from the standard subset to create a new subtest, referred to here as *matched standard*, with approximately the same forced-report baseline performance level as the deceptive subtest: 12 randomly selected items from the standard subset (mean forced-report performance = 26.7%) were combined with 18 randomly selected items from the difficult subset (mean forced-report performance = 0.9%), yielding a forced-report performance level of 11.2% (compared with 11.8% for the deceptive subtest).

A comparison of participants' performance on the matched-standard and deceptive subtests leaves the general conclusions unchanged: Unlike the result for the deceptive subtest, participants were able to increase their free-report accuracy substantially on the matched-standard subtest by more than a factor of five, to 63.0% (based on 27 participants, because 3 participants did not volunteer any of the matched-standard items on the free-report phase). At the same time, free-report quantity performance on the matched-standard subtest decreased slightly to 8.6%. As with the original standard subtest, then, a large accuracy increase was accompanied by a relatively small quantity decrease (both significant). In fact, when comparing the matched-standard and deceptive performance tradeoff patterns, a significantly larger accuracy improvement was evidenced for the matched-standard subtest, $F(1, 25) = 43.19, p < .0001$, but this came at a smaller quantity reduction, $F(1, 28) = 4.34, p < .05$. Thus, if anything, the beneficial consequences of effective monitoring for free-report performance are even more clear when the matched-standard (rather than standard) subtest is considered.

Discussion

The results of Experiment 2 highlight the criticality of monitoring effectiveness for free-report memory performance. When people's confidence judgments are reasonably diagnostic of the correctness of their answers, the option of free report can allow them to achieve high levels of accuracy. In other situations, however, people's monitoring may be undiagnostic to the point of being useless. People still control their memory reporting according to their monitoring output, but the attained level of free-report accuracy may be little better than when they are denied the option of deciding which answers to volunteer.

Of particular importance is the demonstration that monitoring effectiveness can determine memory performance independent of memory "retention." Although retention, as indexed by forced-report quantity performance, was virtually identical for the deceptive and matched-standard subtests, the joint levels of free-report accuracy and quantity performance were far superior for the matched-standard subtest because of the better monitoring associated with this subtest. Clearly, then, free-report memory performance depends on the effective operation of metamemory processes that are simply not tapped by forced-report performance.

This point brings to the fore a basic difference between our proposed conceptual framework and the signal-detection approach to memory. How does the signal-detection framework address the separate contributions of memory retention (or memory strength) and monitoring effectiveness to memory performance? As a matter of fact, it does not. The signal-detection framework does not even allow for a dissociation between subjective confidence and memory strength (Chandler, 1994), and indeed, in that framework confidence is generally taken to index memory strength (see e.g., Lockhart & Murdock, 1970; Parks, 1966). 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 person's confidence distinguishes old-studied from new-foil items) cannot be operationally or conceptually separated: Both are equally valid interpretations of d' (see, e.g., Banks, 1970; 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. Each must be evaluated independently: One may have good monitoring resolution, yet very poor retention, or vice versa. Thus, for instance, poor free-report memory performance might be found to derive from poor retention, poor monitoring, or both.

Several other studies have also recently indicated a dissociation between monitoring and retention. For instance, Kelley and Lindsay (1993) observed 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 (e.g., 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, whereas in fact their actual performance was impaired. She also pointed out that such a finding is contrary to the basic premises of the signal-detection approach to memory.

In sum, by distinguishing the contributions of retention, monitoring, and control, the proposed conceptual framework motivates a greater concern for the operation of the metacognitive processes that mediate free-report memory performance (see also Barnes et al., 1995, Nelson & Narens, 1990, 1994). At the same time, this framework raises the issue of how the separate contribution of each component can be taken into account in the assessment of memory performance (see General Discussion).

General Discussion

This article examined the processes underlying the strategic regulation of memory reporting in free-report memory situations, focusing on one specific means of regulation—the withholding or volunteering of individual items of information. As a framework for investigating such regulation, we put forward a model of monitoring and control processes that merges ideas from signal-detection theory with ideas from metamemory research. The proposed model is necessarily schematic, leaving open such questions as what determines subjective confidence (see, e.g., Gigerenzer, Hoffrage, & Kleinbolting, 1991; Griffin & Tversky, 1992; Kelley & Lindsay, 1993; Koriat, 1993; Metcalfe, Schwartz, & Joaquim, 1993; Miner & Reder, 1994) and precisely how the volunteering criterion is adjusted. Nevertheless, the reported work demonstrates how this model provides a useful analytic tool, specifying patterns of memory accuracy and quantity performance under different testing conditions and bringing forward the critical factors underlying these patterns.

In the following discussion, we first consider the implications of the present work regarding the effects of monitoring and control processes on memory performance, focusing on the moderating role of accuracy motivation and monitoring effectiveness. We then address the general issue of whether and how subject-controlled metamemory processes may be incorporated into the evaluation of memory performance and offer a solution that follows from our framework.

Self-Directed Regulation and Its Performance Consequences

A basic tenet of the proposed theoretical framework is that people can boost the accuracy of their memory reports only by screening out answers that they feel are likely to be incorrect, not by enhancing the overall correctness of their answers. This assumption is supported by the previously reported failures of quantity incentives to enhance overall quantity performance (e.g., Nilsson, 1987; Weiner, 1966a, 1966b). It is also supported by our results here (see also Koriat & Goldsmith, 1994), in which the accuracy advantage of free over forced report was obtained even though only about 1% of the answers differed between the two phases (and the changes were equally likely to be for the better or for the worse). Apparently, then, people cannot readily improve the overall quality of the information that they *retrieve,* but they can improve the quality of what they *report.*

Because accuracy can be enhanced only by screening out answers, that enhancement can be achieved only at the risk of reducing the amount of correct information provided. This potential quantity–accuracy tradeoff has important implications for both participants and experimenters: It requires participants to weigh the relative incentives for providing more accurate versus more complete memory reports when deciding on the most effective control policy for the situation at hand. It also requires experimenters to consider both accuracy and quantity measures in tandem when evaluating free-report memory performance (Klatzky & Erdelyi, 1985; and see later discussion).

Although the fundamental dynamic of a quantity–accuracy tradeoff has been widely acknowledged and studied in the context of forced-report recognition memory (owing primarily to the application of signal-detection theory), in the context of free-report performance it has received much less attention (Klatzky & Erdelyi, 1985). In fact, in the latter context, the quantity–accuracy tradeoff is still an enigma: Both its underlying mechanisms and the conditions affecting its occurrence are poorly understood (see, e.g., Erdelyi et al., 1989; Koriat & Goldsmith, 1994; Roediger et al., 1989). An important contribution of the present work, then, is in showing that neither the accuracy advantage that typically derives from subject control over memory reporting, nor the quantity costs of such control, are inevitable; both were shown to depend on two factors—accuracy incentive and monitoring effectiveness. These two factors have been virtually ignored in traditional, quantity-oriented memory research (which might explain Roediger et al.'s, 1989, observation that a recall-criterion effect on quantity performance is "intuitive, but remarkably little evidence for it exists," p. 255). Accuracy motivation and monitoring effectiveness, however, are critical in determining memory accuracy, and they may have substantial effects on memory quantity performance as well. In what follows, we consider each of these factors in turn.

The Role of Accuracy Motivation

The effects of accuracy motivation on memory performance disclose a general pattern that also emerged in the simulation analyses: Under a modest accuracy incentive, simply giving people the option of free report allows them to achieve a substantial increase in accuracy at a relatively low cost in quantity performance (particularly for polarized recall monitoring; see later discussion). That is, by choosing a fairly low response criterion, people can screen out a relatively large proportion of incorrect answers without screening out many correct answers as well. In contrast, raising the criterion further in response to stronger accuracy incentives improves accuracy even more, but now at a relatively large cost in quantity performance.

This pattern can help resolve some seemingly inconsistent findings in our own as well as other experiments, provided that we take into account the actual volunteering rates (and hence, presumed criterion level) exhibited by the participants. Consider results obtained under a low-to-moderate accuracy incentive. First, when comparing the free- and forced-report conditions under a moderate accuracy incentive in our earlier study (Koriat & Goldsmith, 1994, Experiment 1), accuracy performance increased substantially, with no effect on quantity performance. Similarly, when free-report performance under a low accuracy incentive (8:1 bonus-to-penalty ratio for correct and incorrect answers, respectively) or a moderate accuracy incentive (2:1 bonus-to-penalty ratio) was compared with a no-penalty baseline condition, Barnes et al. (1995) found significant reductions in commission errors, again achieved at insignificant costs in quantity performance. The lack of tradeoff observed in these studies is also consistent with previous recall-criterion research (e.g., Bousfield & Rosner, 1970; Britton et al., 1980; Cofer, 1967; Erdelyi, 1970; Erdelyi et al., 1989; Keppel & Mallory, 1969; Roediger & Payne, 1985; Roediger et al., 1989), indicating that forced recall does *not* generally yield a quantity advantage over free recall (except under special conditions, see later discussion).

Second, however, participants in the moderate-incentive condition of Experiment 1 in the present study apparently set a higher response criterion than did participants in our previous study, and so exhibited a substantial quantity–accuracy tradeoff, even though the payoff schedule and the testing materials were the same in both cases. The different criteria are implied by the differences in volunteering rates for the two studies (.63 for recall and .80 for recognition in the previous study versus .50 for recall and .55 for recognition in Experiment 1 here). Perhaps when the free-report condition follows an initial forced-report phase (as in Experiment 1 here), there is a tendency to be more selective in reporting than when the free-report condition comes first (as in Koriat & Goldsmith, 1994; cf. the observed effect of phase order in Experiment 2 across the deceptive and standard items).

Turning now to the high-incentive conditions, in both Experiment 1 of the present study and Experiment 3 in Koriat and Goldsmith (1994), volunteering rates were relatively low (.40 for recall and .50 for recognition in the latter experiment, and .43 for recall and .47 for recognition here). Consequently, these conditions yielded substantial quantity–accuracy tradeoffs when compared with moderate-incentive conditions (and of course, when compared with forced-report). Furthermore, consistent with the simulation analyses, in each case the increased accuracy was disproportionately more costly in terms of quantity performance than the corresponding increase under the more moderate incentive.

This complex pattern suggests that accuracy motivation must be carefully considered in comparing memory performance across different conditions. By contrast, in previous quantity-oriented studies, quantity incentives had little or no effect on participants' quantity performance (e.g., Nilsson, 1987; Weiner, 1966a, 1966b), leading researchers to the general conclusion that motivation "does not affect memory performance" (Nilsson, 1987, p. 187). Likewise, previous results indicating null or very small effects of recall criterion on memory quantity performance (e.g., Bousfield & Rosner, 1970; Erdelyi et al., 1989; Roediger & Payne, 1985; Roediger et al., 1989) were taken to suggest that criterion effects can generally be ignored (Roediger et al., 1989). This apparent discrepancy underscores the fundamentally different concerns of accuracy-oriented and quantity-oriented memory research, the latter focusing almost exclusively on quantity motivation and quantity performance. Had those other studies, like the experiments here, included a condition with a strong incentive for accuracy, they too would most likely have found accuracy motivation to affect not only accuracy performance but quantity performance as well.

The Role of Monitoring Effectiveness

The second factor that should be of special concern to students of memory accuracy is the effectiveness of people's memory monitoring. Thus, a crucial feature of the proposed framework is its emphasis on the contribution of monitoring effectiveness to memory performance independent of the contributions of "retention" and control. As noted earlier, this distinguishes our approach from the signal-detection framework, in which retention and monitoring effectiveness are combined within a single construct.

Holding retention constant, our results indicate that monitoring effectiveness can have a substantial impact on free-report memory performance. First, as illustrated in the simulation analyses and in Experiment 1, to the extent that people's monitoring output is discriminating of correct and incorrect answers, free-report accuracy should benefit greatly. As monitoring resolution increases, accuracy can be improved with smaller costs in quantity performance, so that at the extreme, with perfect resolution, perfect accuracy might be achieved with no sacrifice in quantity at all (see Figures 3 and 4).

Second, however, both the simulation analyses and the results of Experiment 2 indicate that when monitoring effectiveness is poor, the exercise of strategic control could be primarily or entirely detrimental. For instance, the participants in Experiment 2 exhibited only a minimal ability to monitor the correctness of their answers for the deceptive items and therefore were able to achieve only a minor increase in accuracy (while still sacrificing quantity) given free report. In principle, under more extreme conditions in which resolution is completely lacking, the exercise of control could simply reduce quantity performance with no gain in accuracy at all (see Figure 3, Panel a). Worse yet, there might even be cases in which people's confidence in their answers correlates *negatively* with the answers' likelihood to be correct (see Koriat, 1995a). In that case, the option of free report (and accuracy motivation) could actually be detrimental to memory *accuracy* performance as well!

By distinguishing the separate contributions of retention, monitoring, and control to free-report memory performance, the proposed conceptual framework allows a capitalization on the large body of work that has been carried out on metacognitive processes and their determinants (see, e.g., Gigerenzer et al., 1991; Griffin & Tversky, 1992; Koriat, 1993; Koriat et al., 1980; Lichtenstein et al., 1982; Metcalfe & Shimamura, 1994; Miner & Reder, 1994; Nelson & Narens, 1990, 1994; Schwartz, 1994; Wagenaar, 1988). Such work may prove to be of value in explaining variations in memory performance patterns across different stimulus materials, memory contexts, and participant populations.

As an illustration, consider the puzzle concerning the observation of small but significant recall criterion effects on quantity performance in certain experiments but not in others (see Erdelyi et al., 1989; Roediger et al., 1989). Erdelyi et al. (1989) proposed that the guessing base rate for the stimuli may be the critical factor: When the base rate is relatively high, forced-recall instructions that induce guessing will succeed in increasing the number of correct responses beyond that obtained under standard free-recall instructions. In terms of our framework, this proposal implicates monitoring resolution as the critical factor: A polarized recall monitoring distribution with good resolution (see Figure 4, Panel b) entails a low guessing base rate, and according to the model, little or no quantity–accuracy tradeoff. In contrast, a less polarized monitoring distribution with reduced resolution, reflecting an increased contribution of guessing or plausible inference (e.g., the recognition monitoring in Experiment 1; and see Footnote 11, later), should yield a stronger quantity–accuracy tradeoff. The advantage of casting the "guessability" explanation in terms of metamemory monitoring parameters is that it places the issue within a wider theoretical framework, thus enabling further predictions. According to our model, any condition that causes the distribution of assessed probabilities to be less polarized, or in any other way reduces resolution, should result in a larger quantity–accuracy tradeoff and hence in a greater quantity advantage for forced report. Thus, for instance, Roediger et al. (1989) reported larger recall-criterion effects after a 1 week delay than on an immediate test. If short retention intervals are also associated with better monitoring resolution (see Koriat, 1993), then this finding might be explained within our framework as well.

More generally, recent work on metamemory indicates several factors that can influence metacognitive judgments independent of actual retention (e.g., Chandler, 1994; Kelley & Lindsay, 1993; Reder & Ritter, 1992; Schwartz & Metcalfe, 1992; see Discussion of Experiment 2). Hence, given the mediating role of monitoring and control processes, such factors would be expected to affect free-report performance independent of forced-report performance and to exert differential effects on accuracy-based and quantity-based memory measures. Furthermore, our understanding of memory impairment in certain special populations or following certain experimental interventions (e.g., postevent misinformation) could benefit greatly from an analysis that traces the sources of these impairments to aspects of monitoring and control, as well as, or instead of, retention (e.g., Janowski et al., 1989; Klatzky & Erdelyi, 1985; Koriat et al., 1988; Moscovitch, 1995; Nelson et al., 1990; Perfect & Stollery, 1993; Shimamura & Squire, 1986; Weingardt et al., 1994).

Finally, the lessons of metacognitive research might also be applied to the enhancement of people's memory performance (e.g., eyewitness testimony). Thus, in addition to the traditional techniques designed to improve encoding and retrieval (e.g., Fisher & Geiselman, 1992; Fisher et al., 1989; Herrmann, 1993), we may envisage parallel techniques designed to improve monitoring effectiveness (see Bjork, 1994; Druckman & Bjork, 1994) and engender a more optimal control of memory reporting. Indeed, in pointing out the potential performance consequences of metacognitive judgments in many real-world contexts, Bjork (1994) has stressed that "it is as important to educate subjective experience as it is to educate objective experience" (p. 194; see also Nelson & Narens, 1994). So far, however, efforts to improve memory monitoring (focusing primarily on calibration rather than on resolution) have met with only limited success (e.g., Fischhoff, 1982; Gigerenzer et al., 1991; Koriat et al., 1980; Lichtenstein & Fischhoff, 1980). Further efforts are certainly called for.

Incorporating Subject Control Into the Assessment of Memory Performance

Notwithstanding the advantages just mentioned, the theoretical framework advanced in this article implies a particular view of memory that poses some serious problems for the task of assessing memory. How can we sensibly evaluate a person's memory if memory performance, particularly memory accuracy, is under the person's control? When remembering is seen to involve strategic monitoring and control processes, we must face the question of whether the contributions of such processes can or should be incorporated into memory assessment, or whether they should instead be discounted as extraneous to "true" memory (cf. Klatzky & Erdelyi, 1985; Nelson & Narens, 1994).

The Search for "True" Memory

The aforementioned issue reflects a fundamental dilemma in memory research, which can be illustrated generally with respect to Tulving's (1983) multicomponent model. That model distinguishes 13 elements of the episodic memory system—for instance, the memory *engram,* which is the product of encoding, and *ecphoric information,* which is the product of a process (ecphory) that combines information from both the retrieval cues and the engram. This particular distinction was motivated by results indicating that memory performance depends not only on learning but also on the specific conditions of testing (e.g., Tulving & Thomson, 1973; Watkins & Tulving, 1975). Such findings were taken to "highlight the futility of attempting to make global statements about memory" (Schacter, 1989, p. 691; see also Bjork, 1994). From this perspective, the distinction between "true" memory and extraneous factors becomes problematic: Are we to identify memory with the engram component alone or should we consider, for example, ecphoric processing to be an integral aspect of memory (see Tulving, 1983, p. 180)? (Similar questions can also be raised with regard to other elements in Tulving's model; see Koriat & Goldsmith, 1996a.) It is important to note that this is not just a philosophical issue—one's answer will probably dictate to a large extent his or her ensuing research strategy (see, e.g., Watkins, 1979, 1990).

This same issue may be extended to encompass the relationship between metamemory and memory as well. As pointed out earlier, metamemory research has shown people to call on a variety of strategic decision processes that can affect their ultimate memory performance. For example, at the retrieval stage alone, people can decide whether to initiate search on the basis of their preliminary feeling of knowing (Kolers & Palef, 1976; Nelson & Narens, 1990; Norman, 1973; Reder, 1987, 1988); whether to use an inferential or a direct-retrieval strategy (Reder, 1987, 1988; Reder & Ritter, 1992; Ross, 1989); when to terminate the search (Barnes et al., 1995; Costermans et al., 1992; Gruneberg et al., 1977; Nelson & Narens, 1990; Nelson et al., 1990); and even when the information is reached, whether or not to report it (Barnes et al., 1995; Klatzky & Erdelyi, 1985; Koriat & Goldsmith, 1994) and what level of generality or "grain size" to adopt (Neisser, 1988; Yaniv & Foster, 1995, in press). Thus, here too, the question arises, How should the operation of such strategic-decisional processes be handled in memory research?

Perhaps the most common approach is to treat subject-controlled processing as a nuisance factor that should be eliminated or partialled out in order to achieve a memory measure that has been "cleansed" of extraneous contributions. Thus, Nelson and Narens (1994) noted that

> Ironically, although the self-directed processes are not explicitly acknowledged in most theories of memory, there is an implicit acknowledgment on the part of investigators concerning the importance of such processes. The evidence for this is that investigators go to such great lengths to design experiments that eliminate or hold those self-directed processes constant via experimental control! (p. 8)

Such disparagement of subject control seems to predominate in the quantity-oriented assessment of memory,[9] and curiously enough, this tendency has actually been reinforced by the application of the signal-detection methodology to memory. Of course, signal-detection theory has contributed greatly to an awareness of subject-controlled decision processes in memory. However, because signal-detection measures such as *d'* and the *corrected hit rate* (i.e., hit rate minus false alarm rate; see Swets, 1986) are held to provide an estimate of memory strength that is unbiased by variation in subject control (β, commonly referred to as "response bias"), these measures are often used like other techniques that correct for the effects of guessing (e.g., Budescu & Bar-Hillel, 1993; Cronbach, 1984; Gregg, 1986) in order to obtain a "pure" quantity measure (see also Banks, 1970; Koriat & Goldsmith, 1994). Similarly, with regard to free-report performance, the logic of signal-detection theory has typically motivated researchers to control for criterion effects on memory quantity measures (e.g., by using forced-recall procedures; Erdelyi & Becker, 1974; Klatzky & Erdelyi, 1985) rather than to investigate the contribution of subject-controlled processes to memory accuracy as a topic of interest in its own right.

Clearly, such an approach is unsuitable for the accuracy-oriented evaluation of memory. Consider, for example, the recent upsurge of interest in children's memory, prompted by "a growing concern about children's abilities to provide accurate testimony in legal proceedings" (Ornstein, Gordon, & Baker-Ward, 1992, p. 135; see also Ceci & Bruck, 1993; Ceci, Ross, & Toglia, 1987; Loftus & Davies, 1984; Poole & White, 1991, 1993). With regard to this issue, it is in fact the trustworthiness of what the child decides to report (output-bound accuracy performance) rather than the amount of correct information reported (input-bound quantity performance) that is of foremost concern (though, of course, both aspects may be examined in tandem; see later discussion). Hence, addressing this issue by comparing memory quantity performance across age groups after ensuring a common criterion (by using forced-report testing procedures or calculating a corrected hit rate) would seem to miss the point. Instead, we would like to know, To what extent does the children's combined arsenal of memory and metamemory processes allow them to produce information that can be depended on to be correct? To answer that question, the children's metamemory processes must be allowed to operate and exert their influence (see also Koriat & Goldsmith, 1996a).

In sum, because subject-controlled metamemory processes constitute an important means by which people manage their memory accuracy, one cannot simply circumvent these processes in accuracy-oriented research. On the contrary, when interest centers on the faithfulness of memory, and in particular, on the dependability of memory reports in real-life settings, it would seem imperative to treat the ongoing regulation of memory performance as an intrinsic aspect of memory functioning (see Barnes et al., 1995; Koriat & Goldsmith, 1994, 1996a, 1996c; Metcalfe & Shimamura, 1994; Neisser, 1988; Nelson & Narens, 1990, 1994). An important challenge, then, is to find a way to incorporate the contribution of subject-controlled metamemory processes into the evaluation of memory performance.

Quantity–Accuracy Profiles

The method that we propose for investigating free-report performance involves supplementing the standard point-estimate measures of memory with memory performance profiles that take retention (ecphoric information), monitoring, and control into account. This solution resembles that of plotting memory operating characteristic (MOC) curves in the application of the signal-detection methodology.[10] Like an MOC curve, the proposed *quantity–accuracy profile* (QAP) describes the joint levels of quantity and accuracy performance that can be achieved under given conditions.

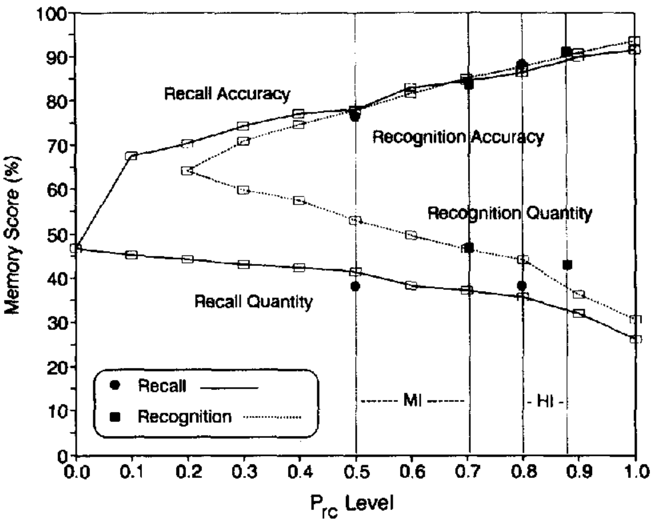

Figure 7. Illustrative quantity–accuracy profiles comparing test format. Potential free-report memory quantity and memory accuracy performance (percent correct) is plotted as a function of P_{rc} level for the recall and recognition participants in Experiment 1 from their forced-report monitoring and performance data. Actual free-report quantity and accuracy measures for each Test Format (recall vs. recognition) × Accuracy Incentive (MI = moderate incentive; HI = high incentive) condition are also plotted as bullets above the P_{rc} estimate for that group. P_{rc} = response criterion probability.

To illustrate the method, Figure 7 presents two QAPs, which were derived at the group level for the recall and recognition participants of Experiment 1. For each participant, the monitoring and performance data from the forced-report phase (i.e., the correctness and assessed probability of each answer in Phase 1) were used to compute the quantity and accuracy performance that would ensue at each response criterion (P_{rc}) level. (The method is essentially the same as that used in the earlier

[9] A notable exception is the treatment of organizational strategies in recall (e.g., Bousfield, 1953; Bower, 1970; Tulving, 1962).

[10] We remind the reader that the signal-detection methodology is generally limited to forced-report recognition tasks. Although Type 2 MOC curves (proportion correct plotted as a function of commission error rate) can be derived from forced recall data (e.g., Murdock, 1966), in that case both *d'* and β lose their usual interpretations: The Type 2 *d'* becomes a measure of monitoring effectiveness (the discrimination of correct from incorrect answers) rather than retention, and "there is no sense in which a Type 2 *d'* 'corrects' recall probabilities for response biases" (Lockhart & Murdock, 1970, p. 108). (For further discussion of the complications inherent in Type 2 analyses, see Banks, 1970; Bernbach, 1967; Healy & Jones, 1973; Murdock, 1974.)

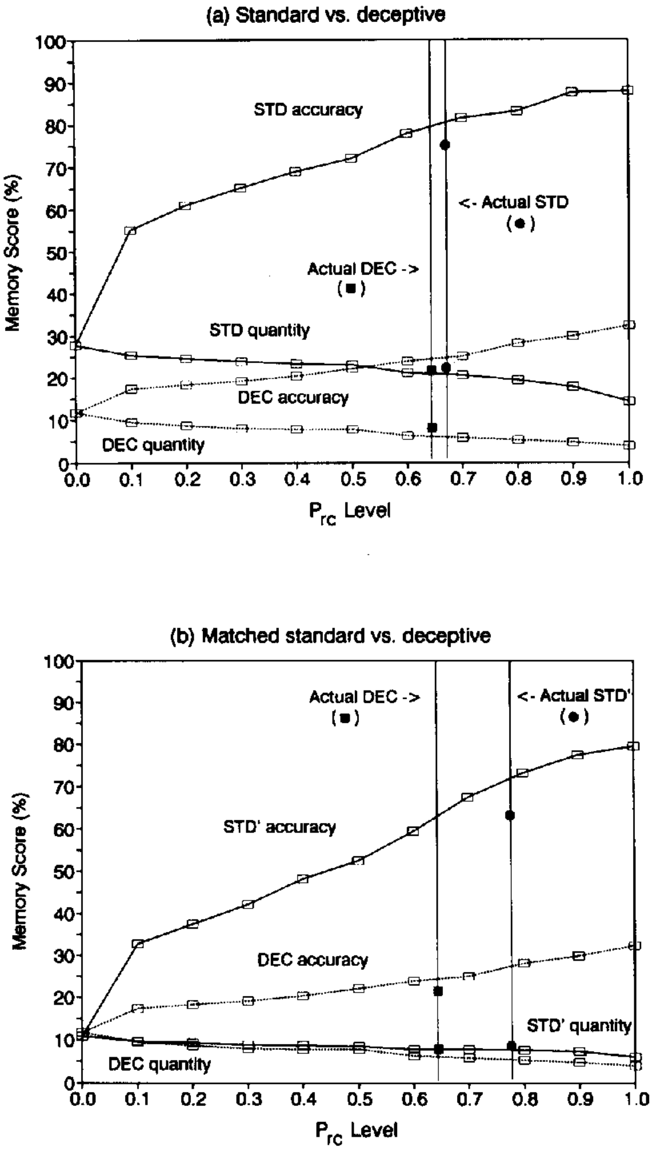

Figure 8. Illustrative quantity–accuracy profiles comparing two levels of monitoring effectiveness. Potential free-report memory quantity and memory accuracy performance (percent correct) is plotted as a function of P_{rc} level for the participants in Experiment 2, comparing the standard (STD) and deceptive (DEC) monitoring conditions (Panel a), and comparing the matched-standard (STD') and deceptive monitoring conditions (Panel b). Actual free-report quantity and accuracy measures for each monitoring condition are also plotted as bullets above the P_{rc} estimate for that condition. P_{rc} = response criterion probability.

simulation analyses, but here the actual monitoring data were used instead of hypothetical data.) The means of the simulated quantity and accuracy scores at each criterion level were then plotted for each test-format group. In addition, the actual means of the free-report accuracy and quantity scores for each experimental group (Test Format × Incentive) are also marked above the mean estimated P_{rc} setting for that group.

A quick comparison of the recall and recognition functions can serve to illustrate some of the potential value that this type of analysis holds as an assessment procedure. Briefly, it can be seen that both of the QAPs exhibit a similar pattern, yet there are some notable differences: First, as may be expected, forced-report performance ($P_{rc} = 0$ for recall, $P_{rc} = .20$ for recognition) is higher for recognition than for recall. Second, however, a sudden jump in accuracy accompanied by little loss in quantity appears for recall participants as soon as they are allowed to withhold any answers at all ($P_{rc} = .10$), whereas the corresponding gain in accuracy for recognition participants at $P_{rc} = .30$ is less pronounced and is accompanied by a symmetric reduction in quantity performance. Third, quantity performance drops more slowly for recall than for recognition as P_{rc} is increased. These latter two trends appear to derive from the greater polarization and resolution of the recall monitoring noted earlier[11] (although they may also stem in part from the different baseline levels of forced-report performance). Overall, however, potential accuracy is as high for recognition as for recall (note also the stricter recognition criterion settings, reported earlier), and thus free-report recognition memory may be generally superior to recall when considering both accuracy and quantity performance together.

Group QAPs could be used to supplement the standard memory measures and provide a more comprehensive picture of the effects of many different kinds of experimental manipulations. Figure 8 presents a more vivid example, in which group QAPs portray the effects of the monitoring manipulation in Experiment 2 here.

QAP assessment may also be applied at the individual level. For illustrative purposes, Figure 9 presents the memory profiles for 6 selected recall participants from Experiment 1 (all of these participants except Participant C were in the high accuracy-incentive condition). If we were to look only at forced-report performance ($P_{rc} = 0$) as a point-estimate of memory retention, we would find clear differences between the participants' performance. However, the profiles offer much more than this. For example, although Participants A and B demonstrated more or less equivalent forced-report performance, B's superior monitoring effectiveness (about twice as high as that of A on the ANDI measure) allows for far greater accuracy potential. Sim-

[11] Although we have included test-format comparisons primarily for illustrative purposes, one could speculate that there may in fact be a systematic difference in the quality of recall and recognition monitoring. Recall testing may tend to induce relatively distinct subjective states of "knowing" and "not knowing," thereby giving rise to a fairly polarized monitoring output. Multiple-choice recognition testing, in contrast, might yield monitoring that is more judgmental in nature, in which participants recruit a variety of considerations to assess the probability that a particular alternative is correct (see Koriat et al., 1980). Thus, we might expect a more graded (less polarized) monitoring output for recognition than for recall, and hence lower monitoring resolution (as was found in Experiment 1). Moreover, to the extent that aspects of the retrieval process itself (e.g., fluency and latency) provide generally valid cues regarding the correctness of candidate answers (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Kelley & Lindsay, 1993; Koriat, 1993; Nelson & Narens, 1990), then being deprived of such cues, recognition monitoring might be expected to be inferior for this reason as well (see Dunlosky & Nelson, in press, mentioned earlier in Footnote 6). This would seem to be an important topic for future research.

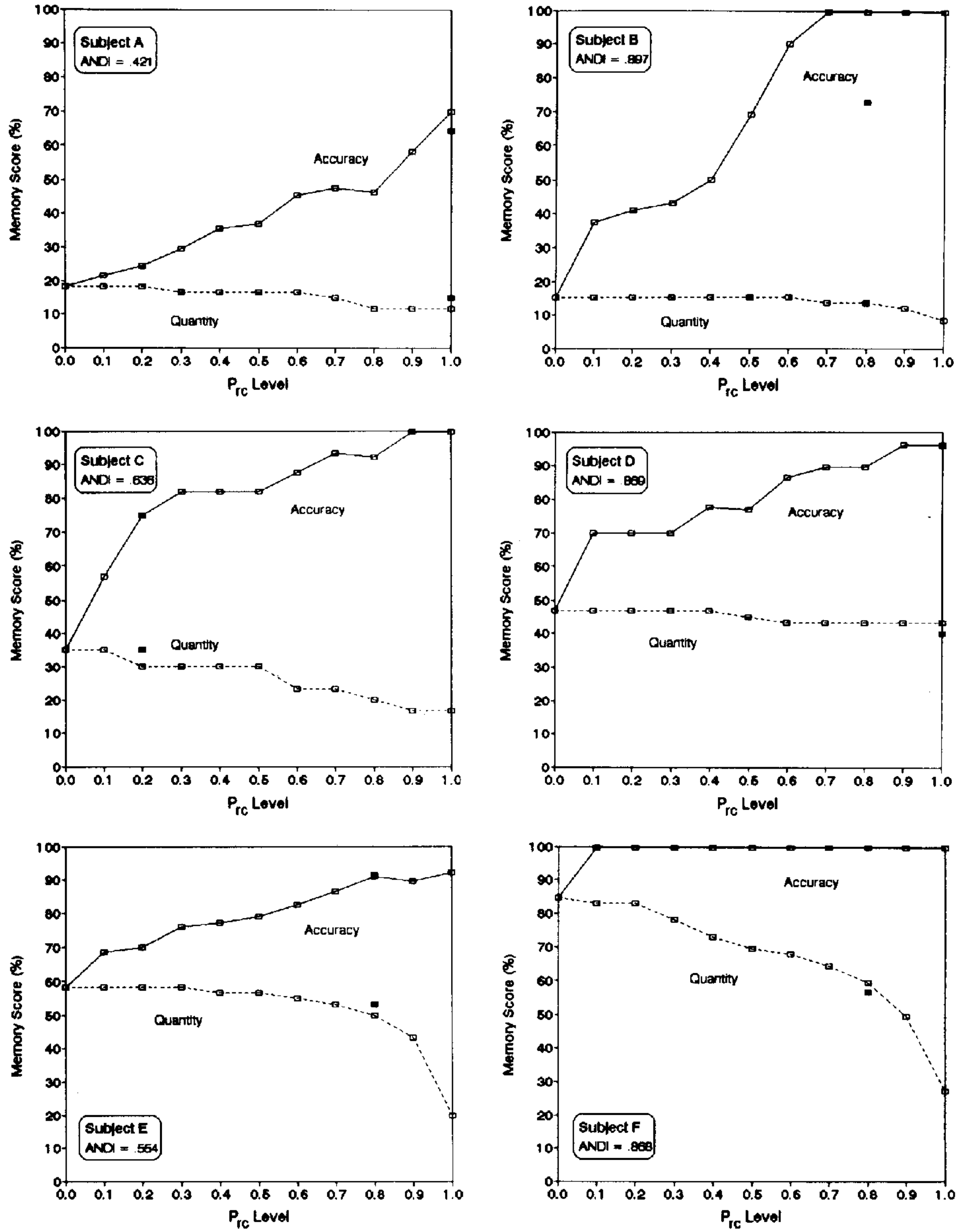

Figure 9. Illustrative quantity–accuracy profiles for six selected recall participants in Experiment 1. In each profile, potential quantity and accuracy performance (percent correct) is plotted as a function of P_{rc} level, based on the forced-report (Phase 1) monitoring and performance data for that participant. Actual free-report (Phase 2) accuracy and quantity scores are also plotted as bullets above the estimated P_{rc} setting used by the participant. Each participant's adjusted normalized discrimination index (ANDI) measure of monitoring resolution (Yaniv et al., 1991) is also presented. P_{rc} = response criterion probability.

ilarly, in comparing Participants C and D, D's potential quantity performance is superior to C's across the range of criterion levels, and D's greater resolution allows a high level of accuracy to be achieved at virtually no cost in the number of correct answers provided. In the next profile, Participant E may achieve a fairly high degree of accuracy with little cost in quantity under a moderate accuracy incentive, yet under a high accuracy incentive will require a large sacrifice in quantity to achieve a relatively small improvement in the dependability of her memory report. Finally, Participant F not only demonstrates excellent memory quantity but also can achieve perfect accuracy when simply given the option of free report. However, this participant is somewhat underconfident about answers that are in fact correct, and therefore the high accuracy incentive induced a substantial proportion of these to be withheld.

According to the QAPs, then, which individuals exhibited the best overall memory? To answer such a question, one may need to take into account functional considerations pertaining to the circumstances of the memory report (see Koriat & Goldsmith, 1996a, 1996c). For instance, in deciding between C's and E's memory performance, despite E's superior quantity potential, for a key witness in a capital trial, we might actually prefer C's memory because of the very high premium placed on memory accuracy in such situations.

QAPs may be used to separate the effects of different variables on memory retention, monitoring, and control in a manner similar to the way signal-detection methods allow one to distinguish differential effects on d' and β. Effects on the retention and accessibility of information can be examined with respect to quantity performance for any given level of confidence (including forced-report). Effects on monitoring can be examined both in terms of resolution indexes, such as gamma (Nelson, 1984) or ANDI (Yaniv et al., 1991), and in terms of potential accuracy across the range of criterion levels. The free-report phase adds information about control: Effects on the control policy (including its optimality; see Footnote 7) can be determined by estimating actual free-report criterion levels using the computational method used here (see Table 2). Such analyses might be applied, for instance, to trace the sources of individual or group differences in memory functioning (e.g., aging, brain damage, etc.) or to investigate the effects of such standard manipulations as retention interval (e.g., Bahrick, Hall, & Dunlosky, 1993) and postevent misinformation (e.g., Weingardt et al., 1994) on both accuracy and quantity performance.

In sum, the focus on memory accuracy calls for an approach in which the operation of metamemory processes is treated as an integral aspect of remembering. The proposed QAP methodology allows the incorporation of monitoring and control processes into the assessment of memory performance, while also affording an evaluation of their separate contributions.

Concluding Remarks

As we noted in introducing this article, remembering in everyday life is guided by a variety of goals other than simply reproducing as much information as possible (see also, Neisser, 1996; Neisser & Fivush, 1994; Winograd, 1996). Toward achieving these goals, rememberers routinely use various means of regulating their memory reporting. They may, for instance, choose what aspects of the event to relate, which to play down or ignore, what perspective to adopt, what level of generality or detail to provide, and so forth (see Neisser, 1981, 1988; Nigro & Neisser, 1983; Ross & Buehler, 1994). Such strategic regulation presents an important challenge to researchers who wish to bring into the laboratory some of the dynamics of remembering in everyday life. Two basic problems emerge (Koriat & Goldsmith, 1996b, 1996c): First, how can subject-controlled regulatory processes be made amenable to experimenter-controlled scientific study? Second, given that memory accuracy is under the strategic control of the rememberer, how can the impact of such control be accommodated by our methods of memory assessment?

In the present work, we took a modest step toward tackling these problems, focusing on one particular type of strategic regulation. Thus, beyond its specific theoretical and empirical contributions, the work illustrates how some of the dynamics underlying the real-life regulation of memory accuracy can be experimentally studied. It also offers a general methodology that can be used to incorporate subject-controlled metamemory processes into the evaluation of memory performance. Finally, the work shows how an accuracy-oriented approach to memory can bring to the fore questions that might be neglected when the focus is strictly on memory quantity. Of course, in everyday life, people have more means available to manage their memory reporting than just the simple option of volunteering or withholding particular items of information. Thus, a better understanding of the strategic regulation of memory performance in real-life contexts will require greater efforts to bring these other aspects of subject control under systematic investigation.

References

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

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

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74,* 81–99.

Barnes, A. E., Nelson, T. O., Dunlosky, J., Mazzoni, G., & Narens, L. (1995). *An integrative system of metamemory components involved in retrieval.* Manuscript submitted for publication.

Bartlett, F. C. (1932). *Remembering.* Cambridge, England: Cambridge University Press.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28,* 610–632.

Bernbach, H. A. (1967). Decision processes in memory. *Psychological Review, 74,* 462–480.

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

Bousfield, W. A. (1953). The occurrence of clustering in recall of randomly arranged associates. *Journal of General Psychology, 49,* 229–273.

Bousfield, W. A., & Rosner, S. R. (1970). Free vs. uninhibited recall. *Psychonomic Science, 20,* 75–76.

Bower, G. H. (1970). Organizational factors in memory. *Cognitive Psychology, 1,* 18–46.

Brewer, W. F. (1996). What is recollective memory? In D. C. Rubin (Ed.), *Remembering our past: Studies in autobiographical memory.* (pp. 19–66). Cambridge, England: Cambridge University Press.

Britton, B. K., Meyer, B. J., Hodge, M. H., & Glynn, S. M. (1980). Effects of organization of text on memory: Tests of retrieval and response criterion hypotheses. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 620–629.

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.). (1976). *Recall and recognition.* London: Wiley.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning & Verbal Behavior, 5,* 325–337.

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.

Ceci, S. J., & Bruck, M. (1993). The suggestibility of the child witness: An historical review and synthesis. *Psychological Bulletin, 113,* 403–439.

Ceci, S. J., Ross, D. F., & Toglia, M. P. (1987). Suggestibility of children's memory: Psycholegal implications. *Journal of Experimental Psychology: General, 116,* 38–49.

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

Cofer, C. N. (1967). Does conceptual organization influence the amount retained in immediate free recall? In B. Kleinmuntz (Ed.), *Concepts and the structure of memory* (pp. 181–214). New York: Wiley.

Cohen, R. L. (1988). Metamemory for words and enacted instructions: Predicting which items will be recalled. *Memory and Cognition, 16,* 452–460.

Conway, M. A. (1991). In defense of everyday memory. *American Psychologist, 46,* 19–27.

Conway, M. A. (1993). Method and meaning in memory research. In G. M. Davies & R. H. Logie (Eds.), *Memory in everyday life* (pp. 499–524). Amsterdam: Elsevier Science Publishers.

Costermans, J., Lories, G., & Ansay, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 142–150.

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

Deffenbacher, K. A. (1991). A maturing of research on the behavior of eyewitnesses. *Applied Cognitive Psychology, 5,* 377–402.

Druckman, D., & Bjork, R. A. (Eds.). (1994). *Learning, remembering, believing: Enhancing human performance.* Washington, DC: National Academy Press.

Dunlosky, J., & Nelson, T. O. (in press). An empirical evaluation of two potential explanations for the delayed-JOL effect: The monitoring-dual-memories explanation versus the transfer-appropriate-monitoring explanation. *Journal of Memory and Language.*

Erdelyi, M. H. (1970). Recovery of unavailable perceptual input. *Cognitive Psychology, 1,* 99–113.

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

Erdelyi, M. H., Finks, J., & Feigin-Pfau, M. B. (1989). The effect of response bias on recall performance, with some observations on processing bias. *Journal of Experimental Psychology: General, 118,* 245–254.

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

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge, England: Cambridge University Press.

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., & Geiselman, R. E. (1992). *Memory enhancing techniques for investigative interviewing: The cognitive interview.* Springfield, IL: Thomas.

Fisher, R. P., Geiselman, R. E., & Amador, M. (1989). Field test of the cognitive interview: Enhancing the recollection of actual victims and witnesses of crime. *Journal of Applied Psychology, 74,* 722–727.

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.

Flanagan, E. J. (1981). Interviewing and interrogation techniques. In J. J. Grau (Ed.), *Criminal and civil investigation handbook* (pp. 4:3–4:23). New York: McGraw-Hill.

Foss, D. J. (Ed.). (1991). [Science Watch special section on everyday memory]. *American Psychologist, 46,* 16–48.

Gardiner, J. M., & Klee, H. (1976). Memory for remembered events: An assessment of output monitoring in free recall. *Journal of Verbal Learning and Verbal Behavior, 15,* 227–233.

Gardiner, J. M., Passmore, C., Herriot, P., & Klee, H. (1977). Memory for remembered events: Effects of response mode and response produced feedback. *Journal of Verbal Learning and Verbal Behavior, 16,* 45–54.

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

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 and psychophysics.* New York: Wiley.

Gregg, V. H. (1986). *Introduction to human memory.* London: Routledge & Kegan Paul.

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

Gruneberg, M. M., Monks, J., & Sykes, R. N. (1977). Some methodological problems with feeling of knowing studies. *Acta Psychologica, 41,* 365–371.

Gruneberg, M. M., & Morris, P. E. (1992). Applying memory research. In M. Gruneberg & P. Morris (Eds.), *Aspects of memory* (2nd ed.; Vol. 1, pp. 1–17). London: Routledge.

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

Herrmann, D. J. (1993). *Improving student memory.* Seattle, WA: Hogrefe & Huber.

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

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, N. L., & Kotz, S. (1970). *Continuous univariate distributions-2.* Boston: Houghton Mifflin.

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.

Keppel, G., & Mallory, W. (1969). Presentation rate and instructions to guess in free recall. *Journal of Experimental Psychology, 79,* 269–275.

Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica, 67,* 95–119.

Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review, 74,* 496–504.

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.

Kolers, P. A., & Palef, S. R. (1976). Knowing not. *Memory and Cognition, 4,* 553–558.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609–639.

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

Koriat, A. (1995b, May). *Illusions of knowing: A window to the link between knowledge and metaknowledge.* Paper presented at the Symposium on Social and Cognitive Aspects of Metacognition, Louvain-la-Neuve, Belgium.

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

Koriat, A., & Goldsmith, M. (1996a). The correspondence metaphor of memory: Right, wrong, or useful? *Behavioral and Brain Sciences, 19,* 211–222.

Koriat, A., & Goldsmith, M. (1996b). Memory as something that can be counted versus memory as something that can be counted on. In D. Hermann, C. McEvoy, C. Hertzog, P. Hertel, & M. Johnson (Eds.), *Basic and applied memory research: Practical applications, Vol. 2* (pp. 3–18). Hillsdale, NJ: Erlbaum.

Koriat, A., & Goldsmith, M. (1996c). Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences, 19,* 167–188.

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

Koriat, A., & Lieblich, I. (1974). What does a person in the "TOT" state know that a person in a "Don't Know" state does not know? *Memory and Cognition, 2,* 647–655.

Koriat, A., & Lieblich, I. (1977). A study of memory pointers. *Acta Psychologica, 41,* 151–164.

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgements: Calibration, discrimination, and monotonicity. *Psychological Bulletin, 114,* 162–173.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26,* 141–171.

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

Loftus, E. F. (1979). *Eyewitness testimony.* Cambridge, MA: Harvard University press.

Loftus, E. F. (1982). Memory and its distortions. In A. G. Kraut (Ed.), *G. Stanley Hall lectures* (pp. 119–154). Washington, DC: American Psychological Association.

Loftus, E. F., & Davies, G. M. (1984). Distortions in the memory of children. *Journal of Social Issues, 40,* 51–67.

Loftus, E. F., & Hoffman, H. G. (1989). Misinformation and memory: The creation of new memories. *Journal of Experimental Psychology: General, 118,* 100–104.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 19–31.

Loftus, G. R., & Wickens, T. D. (1970). Effect of incentive on storage and retrieval processes. *Journal of Experimental Psychology, 85,* 141–147.

May, R. S. (1986). Overconfidence as a result of incomplete and wrong knowledge. In R. W. Scholz (Ed.), *Current issues in West German decision research* (pp. 13–30). Frankfurt am Main, Germany: Lang.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122,* 47–60.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do metamemory ratings affect study time allocation? *Memory and Cognition, 18,* 196–204.

Metcalfe, J. (1993). Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychological Review, 100,* 3–22.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 851–861.

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

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory and Cognition, 15,* 238–246.

Miner, A. C., & Reder, L. M. (1994). A new look at feeling of knowing: Its metacognitive role in regulating question answering. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 47–70). Cambridge: MIT Press.

Moscovitch, M. (1995). Confabulation. In D. Schacter, J. Coyle, G. Fishbach, M.-M. Mesulam, & L. Sullivan (Eds.), *Memory distortion: How minds, brains, and societies reconstruct the past* (pp. 226–251). Cambridge, MA: Harvard University Press.

Murdock, B. B. (1966). The criterion problem in short term memory. *Journal of Experimental Psychology, 72,* 317–324.

Murdock, B. B. (1974). *Human memory: Theory and data.* New York: Wiley.

Murdock, B. B. (1982). Recognition memory. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition.* New York: Academic Press.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12,* 595–600.

Neisser, U. (1981). John Dean's memory: A case study. *Cognition, 9,* 1–22.

Neisser, U. (1988). Time present and time past. In M. M. Gruneberg, P. Morris, & 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.

Neisser, U., & Fivush, R. (Eds.). (1994). *The remembering self: Construction and accuracy in the self-narrative.* New York: Cambridge University Press.

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. (1986). ROC curves and measures of discrimination accuracy: A reply to Swets. *Psychological Bulletin, 100,* 128–132.

Nelson, T. O. (1993). Judgments of learning and the allocation of study time. *Journal of Experimental Psychology: General, 122,* 269–273.

Nelson, T. O., Dunlosky, J., White, D. M., Steinberg, J., Townes, B. D., & Anderson, D. (1990). Cognition and metacognition at extreme altitudes on Mount Everest. *Journal of Experimental Psychology: General, 119,* 367–374.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor in vain" effect. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14,* 476–486.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 125–173). New York: Academic Press.

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

Nigro, G., & Neisser, U. (1983). Point of view in personal memories. *Cognitive Psychology, 15,* 467–482.

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

Norman, D. A. (1973). Memory, knowledge, and the answers of questions. In R. L. Solso (Ed.), *Contemporary issues in cognitive psychology: The Loyola Symposium* (pp. 135–165). Washington, DC: Winston.

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.

Ornstein, P. A., Gordon, B. N., & Baker-Ward, L. (1992). Children's memory for salient events: Implications for testimony. In M. L. Howe, C. H. Brainard, & V. F. Reyna (Eds.), *Development of long-term retention* (pp. 135–158). New York: Springer-Verlag.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs, 76,* (28, Whole No. 547).

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

Perfect, T. J., & Stollery, B. (1993). Memory and metamemory performance in older adults: One deficit or two? *Quarterly Journal of Experimental Psychology, 46*A, 119–135.

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: Effects of question repetition and retention interval on the eyewitness testimony of children and adults. *Developmental Psychology, 29,* 844–853.

Pressley, M., Levin, J. R., Ghatala, E. S., & Ahmad, M. (1987). Test monitoring in young grade school children. *Journal of Experimental Child Psychology, 43,* 96–111.

Puff, C. R. (Ed.). (1982). *Handbook of research methods in human memory and cognition.* New York: Academic Press.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology, 19,* 990–138.

Reder, L. M. (1988). Strategic control of retrieval strategies. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 227–259). New York: Academic Press.

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.

Roediger, H. L., & Payne, D. G. (1985). Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories. *Memory and Cognition, 13,* 1–7.

Roediger, H. L., Srinivas, K., & Waddil, P. (1989). How much does guessing influence recall? Comment on Erdelyi, Finks, and Feigin-Pfau. *Journal of Experimental Psychology: General, 118,* 253–257.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96,* 341–357.

Ross, M. (in press). Validating memories. In N. L. Stein, P. A. Ornstein, B. Tversky, & C. Brainerd (Eds.), *Memory for everyday and emotional events.* Hillsdale, NJ: Erlbaum.

Ross, M., & Buehler, R. (1994). Creative remembering. In U. Neisser & R. Fivush (Eds.), *The remembering self: Construction and accuracy in the self-narrative* (pp. 205–235). New York: Cambridge University Press.

Schacter, D. (1989). Memory. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 687–725). Cambridge, MA: MIT Press.

Schacter, D. (1995). Memory distortion: History and current status. In D. Schacter, J. Coyle, G. Fishbach, M.-M. Mesulam, & L. Sullivan (Eds.), *Memory distortion: How minds, brains, and societies reconstruct the past* (pp. 1–43). Cambridge, MA: Harvard University Press.

Schacter, D. L., Coyle, J. T., Fishbach, G. D., Mesulam, M.-M., & Sullivan, L. E. (Eds.). (1995). *Memory distortion: How minds, brains, and societies reconstruct the past.* Cambridge, MA: Harvard University Press.

Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology, 9,* 321–332.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin and 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.

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 93–113). Cambridge: MIT Press.

Shepard, R. N. (1967). Recognition memory for words, sentences and pictures, *Journal of Verbal Learning and Verbal Behavior, 6,* 156–163.

Shimamura, A. P., & Squire, L. R. (1986). Memory and metamemory: A study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 452–460.

Stern, W. (1904). Wirklichkeitsversuche. [Realistic experiments]. *Beitrage zur Psychologie der Aussage, 2,* 1–31.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

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

Tulving, E. (1962). Subjective organization in free recall of "unrelated" words. *Psychological Review, 69,* 344–354.

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

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

Wagenaar, W. A. (1988). Calibration and the effects of knowledge and reconstruction in retrieval from memory. *Cognition, 28,* 277–296.

Watkins, M. J. (1979). Engrams as cuegrams and forgetting as cue overload: A cueing approach to the structure of memory. In C. R. Puff

(Ed.), *Memory organization and structure* (pp. 347–372). New York: Academic Press.

Watkins, M. J. (1990). Mediationism and the obfuscation of memory. *American Psychologist, 45,* 328–335.

Watkins, M. J., & Tulving, E. (1975). Episodic memory: When recognition fails. *Journal of Experimental Psychology: General, 104,* 5–29.

Weiner, B. (1966a). Effects of motivation on the availability and retrieval of memory traces. *Psychological Bulletin, 65,* 24–37.

Weiner, B. (1966b). Motivation and memory. *Psychological Monographs: General and Applied, 80,* (18, Whole No. 626).

Weingardt, K. R., Leonesio, R. J., & Loftus, E. F. (1994). Viewing eyewitness research from a metacognitive perspective. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 157–184). Cambridge: MIT Press.

Wells, G. L., & Lindsay, R. C. L. (1985). Methodological notes on the accuracy–confidence relation in eyewitness identifications. *Journal of Applied Psychology, 70,* 413–419.

Wells, G. L., & Loftus, E. F. (Eds.). (1984). *Eyewitness testimony: Psychological perspectives.* Cambridge, England: Cambridge University Press.

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

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

Winograd, E., & Neisser, U. (Eds.). (1992). *Affect and accuracy in recall: Studies of "flashbulb memories."* New York: Cambridge University Press.

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. (in press). Precision and accuracy in judgmental estimation. *Journal of Behavioral Decision Making.*

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

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30,* 132–156.

Yates, J. F. (1990). *Judgment and decision-making.* Englewood Cliffs, NJ: Prentice-Hall.

Appendix

Details of the Simulated Monitoring Manipulations

The simulation analyses exploring the effects of monitoring on free-report memory performance assumed a situation in which candidate answers are assigned to one of 11 assessed-probability categories (0, .10, . . . , .90, 1.0) and exactly 50% of the answers are correct. The quality of monitoring was manipulated by independently varying the proportion-correct for each category (*correspondence*), and the relative-frequency of each category (*polarization*). Each manipulation will now be detailed.

The correspondence between the proportion correct and the assessed probability of each category was manipulated using the parameter s in the following formula:

$$PC_i = PA_i + [(1 - s)(0.5 - PA_i)] \qquad [s: 0 \leq s \leq 1.0], \quad \text{(A1)}$$

where for each category, i, PC_i is the proportion correct and PA_i is the assessed-probability level.

It may be seen that s is the slope of the resulting calibration function. When $s = 0$, then $PC_i = .50$ for all categories (no correspondence), and when $s = 1$, then $PC_i = PA_i$ for all categories (perfect correspondence). Intermediate values of s determine the extent to which PC_i approaches its perfectly calibrated value of PA_i.

The polarization manipulation varied the distribution of the candidate answers among the 11 assessed-probability levels (RP_i—relative proportion in each category) according to the parameter d. We first defined three boundary distributions: $d = -1.0$, all answers are assigned to the category PA = .50 (unipolar); $d = +1.0$, half of the answers are assigned to each of two categories, PA = 0 and PA = 1.0 (bipolar); $d = 0$, the answers are evenly distributed between the 11 categories, that is, $RP_i = 1/11$ (uniform).

Between these boundary distributions, d determined the shape parameters, p and q, of the standard-form beta distribution (Johnson & Kotz, 1970, p. 37), which was used to set the RP of each category as follows:

(a) The shape parameters were set according to the formula:

$$p = q = 1.8^{(-d)(10)} \quad [d: -1.0 < d < 0]$$

$$p = q = 1.5^{(-d)(10)} \quad [d: 0 < d < 1.0]. \qquad \text{(A2)}$$

(b) The cumulative proportion (CP_i) of answers at or below each assessed-probability level (i), indexed from lowest ($PA_1 = 0$) to highest ($PA_{11} = 1.0$), was then determined by the function:

$$CP_i = \mathrm{PROBBETA}(i/11, p, q). \qquad \text{(A3)}$$

The function returns the probability value of the standard-form beta distribution with parameters p and q, evaluated at the point, $i/11$.

(c) The relative proportion (RP_i) for each category was computed from the obtained cumulative distribution.

Equation A2 was designed to achieve a gradual bipolarization ($d > 0$) and unipolarization ($d < 0$) of the distribution across the range of d values. Because $p = q$, the distribution of answers is always symmetric about the $PA = .50$ category.

Also note that both the correspondence and the polarization manipulations leave the overall proportion of correct answers unchanged ($PC_{\text{total}} = .50$).

Received December 9, 1994
Revision received January 24, 1996
Accepted January 25, 1996 ■