Journal of Experimental Psychology:
Learning, Memory, and Cognition
2008, Vol. 34, No. 5, 1224–1245

0278-7393/08/$12.00 DOI: 10.1037/a0012938

Control Over Grain Size in Memory Reporting—With and Without Satisficing Knowledge

Rakefet Ackerman and Morris Goldsmith
University of Haifa

When answering questions from memory, respondents strategically control the precision or coarseness of their answers. This grain control process is guided by 2 countervailing aims: to be informative and to be correct. Previously, M. Goldsmith, A. Koriat, and A. Weinberg Eliezer (2002) proposed a *satisficing model* in which respondents provide the most precise answer that passes a minimum-confidence report criterion. Pointing to social-pragmatic considerations, the present research shows the need to incorporate a minimum-informativeness criterion as well. Unlike its predecessor, the revised, "dual-criterion" model implies a distinction between 2 theoretical knowledge states: Under moderate-to-high levels of *satisficing knowledge,* a grain size can be found that jointly satisfies both criteria—confidence and informativeness. In contrast, under lower levels of *unsatisficing knowledge*, the 2 criteria conflict—one cannot be satisfied without violating the other. In support of the model, respondents often violated the confidence criterion in deference to the informativeness criterion, particularly when answering under low knowledge, despite having full control over grain size. Results also suggest a key role for the "don't know" response, which when available, can be used preferentially to circumvent the criterion conflict.

Keywords: memory accuracy and informativeness, metamemory, monitoring and control, don't know, gist memory

This article deals with the process of answering questions from memory when the respondent has the option of providing the answer at different levels of *granularity*—precision or coarseness. For example, when asked "How old was [ex-U.S. president] John F. Kennedy at the time of his assassination?", a person might respond with a relatively precise answer, such as "46," or a coarser answer, such as "in his mid 40s," "between 40 and 60," or "quite young." Which of the potential answers will a person actually volunteer, and on what will this depend?

There are several different approaches one might take to address this question. Of course, the granularity of the responses that are produced from memory should depend, at least in part, on the operation of memory encoding, retrieval, and reconstruction processes that have traditionally been the target of study in memory research (e.g., Brainerd & Reyna, 1990; Gernsbacher, 1985; Kintsch, Kozminsky, Streby, McKoon, & Keenan, 1975; Murphy & Shapiro, 1994). In fact, it is often assumed that the grain size of one's answers simply reflects the level of detail of the information that can be accessed. Research on this topic (e.g., gist vs. verbatim memory) has focused mainly on issues concerning memory representation, raising questions such as the following: Are the representations hierarchal or associative, how are the representations processed, and how does the accessibility of the represented information at different levels change over time (Dorfman & Mandler, 1994; Kintsch, Welsch, Schmalhofer, & Zimny, 1990; Koriat, Levy-Sadot, Edry, & de Marcas, 2003; Reyna & Brainerd, 1995; Reyna & Kiernan, 1994)? These aspects of the question can be addressed in terms of the traditional notions of memory representation and processing per se. Thus, for example, the finding that memory for gist is more stable than memory for detail (Kintsch et al., 1990) could be taken to imply that one's answers will become more coarse at longer retention intervals (see Goldsmith, Koriat, & Pansky, 2005). Of course, a variety of other factors besides passage of time can also affect the relative accessibility of coarse versus precise information (e.g., Brainerd & Reyna, 1990; Reyna & Brainerd, 1995).

A very different set of factors, however, is emphasized by the *social-communication approach*, which focuses on cooperative pragmatic principles involved in the explicit and implicit communication that takes place when one person asks a question of another person in a particular social context. Such research has shown people's responses to questions to take into account pragmatic considerations and tacit assumptions relating to the background and existing knowledge of the questioner, his or her purpose in asking the question, personal goals, self expectations, and so forth (e.g., Clark & Schober, 1992; Gibbs & Bryant, 2008; Hilton, 1995; Schwarz, 1999; Sperber & Wilson, 1995). According to Grice's (1975) maxim of quantity, for

Rakefet Ackerman and Morris Goldsmith, Department of Psychology, University of Haifa, Haifa, Israel.

We gratefully acknowledge support of this project by the German Federal Ministry of Education and Research (BMBF) within the framework of German–Israeli Project Cooperation (DIP). Facilities for conducting the research were provided by the Institute of Information Processing and Decision Making, University of Haifa, and by the Max Wertheimer Minerva Center for Cognitive Processes and Human Performance. We thank Asher Koriat, Ainat Pansky, Ruth Kimchi, and Ilan Yaniv for valuable conversations and comments. We also thank Elinor Rosen for her help in running the experiments. The article is based on parts of a doctoral dissertation submitted to the University of Haifa by Rakefet Ackerman.

Correspondence concerning this article should be addressed to Rakefet Ackerman or Morris Goldsmith, Department of Psychology, University of Haifa, Haifa 31905, Israel. E-mail: rakefet@research.haifa.ac.il or mgold@research.haifa.ac.il

example, speakers are expected to make their contribution as informative as required in a particular context but no more so. In line with this principle, participants have been found to adjust the detail of the information they convey according to their perception of how much the listener needs to know (Gibbs & Bryant, 2008; Isaacs & Clark, 1987; van der Henst, Carles, & Sperber, 2002; Vandierendonck & Van Damme, 1988). Highlighting some further social-communicative aspects of memory reporting, participants have been found to focus more on story details and narrative structure in recalling a story to an experimenter than when conveying it to a peer (Hyman, 1994), to include fewer details and verbatim quotes in recounting events when the goal was to entertain than when accuracy was emphasized (Dudukovic, Marsh, & Tversky, 2004; Wade & Clark, 1993), and to convey less detailed information to inattentive than to attentive listeners (Pasupathi, Stallworth, & Murdoch, 1998).

Finally, a third approach still is the *metacognitive* approach. This approach shares the emphasis of the social-communicative approach on strategic behavior guided by personal and social goals, while at the same time attempting to specify the monitoring and control processes that underlie strategic memory performance, and integrate these processes into memory theory (e.g., Barnes, Nelson, Dunlosky, Mazzoni, & Narens, 1999; Benjamin, 2008; Goldsmith & Koriat, 1999, 2008; Koriat & Goldsmith, 1994, 1996b; Koriat, Goldsmith, & Halamish, 2008; Nelson & Narens, 1990). This approach, then, can potentially provide a bridge between the preceding two approaches. Indeed, metacognitive research focusing on the strategic regulation of memory reporting (for a review, see Goldsmith & Koriat, 2008) has shown how memory performance—its quantity and accuracy—depends on an interaction between the contents of memory and metacognitive processes that guide memory search and retrieval, convert retrieved information into concrete candidate answers, determine whether the answers are reported, and in what manner. Such work has also highlighted the need to consider how these monitoring and control processes, and hence memory performance, may be affected by social-functional variables that are characteristic of remembering in real-life situations (Koriat & Goldsmith, 1996a). This line of work, with its underlying approach, is carried forward in this article.

In the present study, we are interested in how people regulate the grain size (precision or coarseness) of answers to memory questions, such as those that might be posed in the course of one's job or in a casual conversation (e.g., "What were our third quarter earnings last year?" or "What is the flying distance between London and Paris?"). Previous work (Goldsmith & Koriat, 1999; Goldsmith et al., 2005; Goldsmith, Koriat, & Weinberg Eliezer, 2002; see also Yaniv & Foster, 1995, 1997) has shown the control of grain size in memory reporting to be guided by two main objectives: *correctness* (cf. Grice's, 1975, maxim of quality) and *informativeness* (cf. Grice's, 1975, maxim of quantity).[1] Importantly, under conditions in which one is uncertain about the correctness of one's memory, these two goals tend to conflict: To increase the chance that one's answer is correct, one should increase the coarseness of the answer (e.g., "300–400 km" rather than "340 km"). Coarsening one's answer generally increases the likelihood that it is correct, both because it increases the number of possible true values that are consonant with the answer, and because memory for coarse information or "gist" is typically available even when memory for precise details is missing (e.g., Brainerd & Reyna, 1990; Friedman, 1979; Kintsch et al., 1990). By doing so, however, the answer becomes less informative and, hence, will generally be less appreciated by the recipient of the information (Yaniv & Foster, 1995).

Given this trade-off dynamic, how do respondents find an appropriate compromise between correctness and informativeness in choosing a grain size for their answers? The grain-control model proposed and examined so far (Goldsmith et al., 2002) emphasizes the aim of correctness over the aim of informativeness, incorporating the goal of correctness by way of an explicit confidence criterion, whereas the goal of informativeness is handled in a more indirect manner. In this article, we put forward a revised model that incorporates an explicit informativeness criterion as well. Focusing on social-pragmatic factors and some existing results, our analysis suggests that although the original single-criterion model may be adequate under conditions in which respondents have a moderate-to-high level of knowledge on which to base their answers, a dual-criterion model is needed to explain the choice of grain size when knowledge is relatively poor. This analysis, which provides the rationale for the development of the revised model, will now be presented.

The Satisficing Model for the Control of Grain Size

Goldsmith et al. (2002) put forward a decision-theoretic, metacognitive model of the control of grain size in memory reporting that takes into account the correctness and informativeness considerations. According to their *satisficing model* (cf. Simon, 1956), the respondent strives to provide as much information as possible, as long as its subjective probability of being correct satisfies some reasonable minimum level. Thus, for example, one might try to specify one's answer to the nearest year, to the nearest 5 years, 10 years, and so forth, until one believes that it has, say, at least an 80% chance of being correct. Only then will the answer be reported. The level of this minimum-confidence criterion (80% in the preceding example) is assumed to depend on the relative incentives for correctness and informativeness in answering a particular question in a particular social situation: A higher criterion is set when correctness is emphasized, a lower criterion is adopted when informativeness is emphasized.

As an alternative to this simple satisficing model, Goldsmith et al. (2002) also examined a more complex, *relative expected-utility maximizing model*. In this model, respondents calculate the subjective expected utility of candidate answers at various grain sizes (taking into account both the subjective value of a correct or incorrect answer, and the subjective probability that the answer is correct or incorrect), compare these values, and ultimately choose the answer whose subjective expected utility is maximal. Such a relative comparison process, while aiming for a more optimal grain-choice solution than

[1] In previous work, the two goals were referred to as *accuracy* and *informativeness*. Though a bit awkward, in this article we use the term *correctness* to avoid ambiguities involving the use of the term accuracy as a synonym for "precision" or as denoting the graded distance between a provided answer and the actual value (cf. Yaniv & Foster, 1995, 1997). Following Goldsmith et al. (2002, 2005), we use the term correctness in a dichotomous sense at the level of individual answers: An answer is *correct* if it contains the true value and is *incorrect* otherwise. Nevertheless, in keeping with common usage, the term accuracy is used in referring to the correctness of a *set* of answers, that is, to denote the proportion of one's answers that are correct.

the satisficing model, seemingly places a much heavier cognitive and metacognitive burden on the rememberer.

In a series of three experiments, Goldsmith et al. (2002) had participants answer a set of general knowledge questions, each soliciting an item of quantitative information: time, date, age, distance, speed, and so forth. Grain size could then be operationalized in terms of the interval width of the provided answer (e.g., "1963," "1960–1970," and "1950–2000" representing answers of increasing coarseness; see also Goldsmith et al., 2005; Yaniv & Foster, 1995, 1997). The questions were presented in two phases. In the first phase, participants answered each item using two different bounded intervals specified by the experimenter (in some cases, the fine-grained answer was elicited as a specific value, e.g., "1963"; an interval width of one). For example, "When did Boris Becker last win the Wimbledon men's tennis finals? (A) Provide a 3-year interval; (B) Provide a 10-year interval." In two of the experiments (Experiments 2 and 3), the participants also rated their confidence in the answer at each grain size by assessing its probability of being correct (i.e., of including the correct value). Immediately following this initial phase, the participants went over their answers, and for each item indicated which of the two grain sizes they would prefer to provide, assuming that they were "an expert witness testifying before a government committee" (Experiments 1 and 2) or in order to earn monetary payoffs, which were larger for correct fine-grained answers than for correct coarse-grained answers (Experiment 3).

On the whole, the results support the simple satisficing model: Across the three experiments, in the critical second phase participants chose to provide the fine-grained answer for about 55% of the items and the coarse-grained answer for the other 45%, implying that the choice of grain size was guided neither solely by the desire to be correct nor solely by the desire to be informative. Instead, the participants tended to provide the coarse-grained answer when the fine-grained answer had a low subjective probability of being correct. Importantly, the choice of grain size was shown to be strategic, taking into account the relative incentives for informativeness versus correctness (Experiment 3): Under a relatively high payoff for informativeness, a lower, more liberal confidence criterion was adopted, and hence more fine-grained answers were provided relative to a payoff scheme that placed a higher emphasis on correctness over informativeness.

Additional analyses indicated that the grain control decision was based primarily on confidence in the fine-grained answer, as predicted by the satisficing model, and not on the relative disparity between confidence in the precise- and coarse-grained answers, as predicted by the relative expected-utility maximizing model. Moreover, modeling the grain choice in terms of a simple criterion on fine-grained confidence successfully accounted for about 90% of the participants' actual grain choices—a significantly better fit than the one achieved by the relative expected-utility maximizing model (76%).

Does the Satisficing Model Satisfice?

Notwithstanding the success of the simple satisficing model in accounting for the basic pattern of results observed so far (see also Goldsmith et al., 2005; Weber & Brewer, 2008), this pattern has been based on a rather restricted experimental paradigm that deviates from the real-life control of grain size in various ways. For one thing, real-life rememberers are not confined to just two possible grain sizes, specified in advance by an experimenter. Instead, in principle, they have unlimited control over the grain size of their answers and, hence, can choose to provide as coarse an answer as is needed to reach a desired level of confidence. Can the satisficing model successfully predict performance under such conditions?

As described above, according to the satisficing model, in choosing a grain size for their answers, respondents protect the goal of correctness explicitly, by setting and adhering to a minimum-confidence criterion, whereas the goal of informativeness is handled implicitly, by providing the most precise answer that passes the confidence criterion. This introduces an inherent asymmetry into the model: Although in some situations one might lower the confidence criterion to give more weight to informativeness (e.g., when trying to impress a new acquaintance with one's knowledge), once the confidence criterion has been set, it constitutes the sole hard constraint on the answering process. Consequently, if during the course of trying to answer a question, one realizes that one's knowledge is so poor that one can only pass the confidence criterion by providing a ridiculously coarse and uninformative answer, one has no choice but to do so—there is no constraint that ensures that respondents' answers are reasonably informative. Thus, for example, when asked in the course of a conversation "When was the Disney movie *Snow White and the Seven Dwarfs* first released?", if the respondent has no idea, he or she may have to produce a ridiculously coarse answer such as "sometime between 1880 and 1970" to be reasonably sure (e.g., 80%) that the answer is correct. According to the satisficing model, this is what the person will do. But is it?

There are reasons to believe that it is not. Some suggestive evidence comes from a study by Yaniv and Foster (1997), who compared several different methods of eliciting interval-type estimates for quantitative information. Using one such method (Study 2), the participants were instructed to provide interval answers to general-knowledge questions as wide as needed to ensure that 95% of their answers would be correct (i.e., would include the true value). In terms of Goldsmith et al.'s (2002) satisficing model, the participants were essentially instructed to set a very high (~95%) confidence criterion in this task. Yet, Yaniv and Foster observed that only 47% of the participants' answers were correct, and in fact, the provided intervals would need to be widened across the board by a factor of 17 to achieve the specified hit rate of 95%. Similar results were found using the other two elicitation methods (see also Klayman, Soll, González-Vallejo, & Barlas, 1999; Soll & Klayman, 2004). Why did the participants so exceedingly fail to coarsen their answers to the extent called for by the task?

One possible reason is overconfidence (Lichtenstein, Fischhoff, & Phillips, 1982; Soll & Klayman, 2004): Yaniv and Foster's (1997) participants may have believed that their answers were coarse enough to achieve a very high accuracy rate, though in fact they were not. Yet, in light of the magnitude of the overconfidence that one would have to assume to account for their results, Yaniv and Foster suggested the contribution of an additional factor: Participants may be reluctant to provide extremely coarse answers that violate social norms of communication, specifically, the expectation that one's answers should be reasonably informative (Grice, 1975). Yaniv and Foster observed, for example, that the strategy of providing enormously broad estimates (e.g., "zero to one billion") in response to 95% of the items would be adaptive in their 95%-confidence-interval study but not in real life,

where such estimates would be subject to ridicule. Thus, the influence of pragmatic communication norms that constrain the coarseness of socially acceptable answers "may diverge from and even supersede the demands of calibration accuracy" (Yaniv & Foster, 1997, p. 30).

Although direct evidence for this idea was lacking in Yaniv and Foster's (1997) study, it has several potentially important implications. First, with regard to the study of calibration of confidence in interval estimation (e.g., Soll & Klayman, 2004), it implies that some of the observed miscalibration may stem from factors related to control, that is, from unwillingness to provide extremely uninformative estimates, rather than from factors related to the accuracy of monitoring. Hence, to the extent that such a control bias exists, its effects might be mistakenly attributed to monitoring-judgment bias. Second, with regard to the study of the control of grain size in memory reporting, it implies that some mechanism must exist for ensuring that one's answers do not violate communication norms requiring a minimum level of informativeness. The simple satisficing model considered so far lacks such a mechanism. In what follows, we propose a revised, dual-criterion model that incorporates such a mechanism, and we outline some predictions that can be derived from it. These predictions are then examined in several experiments.

Minimum Informativeness Criterion

We noted earlier that the original satisficing model is asymmetric, protecting the goal of correctness with an explicit confidence criterion while handling the goal of informativeness in an implicit manner. We now propose to remove this asymmetry by including an explicit *minimum-informativeness* criterion to supplement the confidence criterion. Thus, in the revised, *dual-criterion model*, respondents provide the most precise candidate answer that passes both the confidence criterion and the informativeness criterion. The informativeness criterion reflects the minimum level of precision, or maximum level of coarseness, that is perceived as constituting a reasonably informative answer to the question. Beyond this level, the answer would be perceived as unacceptably coarse (Grice, 1975),[2] and the respondent might even be perceived as responding cynically (Sperber & Wilson, 1995).

Of course, what is considered to be a reasonable level of precision for one's answer can be highly dependent on the particular question that is being asked. For example, consider the following: "What is the population of New York?" versus "How many players are there on a hockey team?" An interval width of 100 would be extremely (overly) precise in answering the first question but unacceptably coarse in answering the second question. More generally, the minimum-informativeness criterion should depend on such factors as the perceived expectations and needs of the questioner in asking a particular question (e.g., how much he or she already knows about the topic; the use to which the information will be put), the respondent's own self-expectations concerning the question (e.g., whether it is in his or her domain of expertise), and social norms for particular types of questions in particular contexts (e.g., when asked "when were you born?" in casual conversation vs. when filling out insurance forms). Consequently, we assume that the informativeness criterion is set (at least implicitly) after each question has been presented but before the answering process begins.

To help conceptualize the implications of the proposed addition of a minimum-informativeness criterion, and understand how the revised, dual-criterion satisficing model differs from the original one, we now draw a distinction between two theoretical knowledge states that fall out of the new model: *satisficing knowledge* (SK) and *unsatisficing knowledge* (UK). One is in an SK state when one's level of knowledge is sufficient to allow one to provide an answer that simultaneously satisfies both the confidence and the informativeness criteria. This state is depicted schematically in a "confidence–informativeness trade-off diagram" (see Figure 1A), which illustrates how changes in the interval width of one's answer (the horizontal widths of the V interior) inversely affect the confidence in and informativeness of the answer. In this example, assuming that one's level of confidence in the correctness of a very precise answer (Answer *A*, confidence = 10%) is below the confidence criterion (65%), the criterion level can nevertheless be reached by coarsening the answer to include a wider range of values (Answer *B*). The grain of this answer is still informative enough to pass the informativeness criterion (i.e., it is not "unacceptably coarse"; cf. Answer *D*). Of course, confidence could perhaps be increased even further by providing an even coarser answer that still passes the informativeness criterion (i.e., any other answer in the range of criterion overlap). Nevertheless, the revised model continues to assume, in accordance with the single-criterion satisficing model, that one aims to provide the most precise answer that passes the confidence criterion (*B* rather than *C* or similar answer).[3]

Note that when respondents have SK, there is essentially no difference between the dual-criterion model and the original satisficing model. The added informativeness criterion has no role in affecting the course of the grain control process or the grain size of the answer that is ultimately chosen. In fact, one way of conceptualizing the original model is to say that it (implicitly) assumed that respondents always have SK. The revised model, however, allows a further possibility, in which respondents are unable to simultaneously satisfy both the confidence and informativeness criteria. This is the UK state, depicted schematically in Figure 1B. Note that this diagram assumes the same situational context and, hence, maintains the same minimum-confidence criterion (i.e., 65%) and minimum-informativeness criterion (at the same answer–interval width)

[2] At this stage, we make no claims regarding the specific nature of the informativeness criterion and its underlying dimension. For example, the criterion might be set in information-theoretic terms (e.g., a *minimum reduction in uncertainty*), in which case the grain size of a particular candidate answer would have to be analyzed and transformed into information units before being compared against the criterion. Alternatively, the criterion might be set directly in terms of a *minimum acceptable precision* (e.g., interval width or taxonomic-categorical level), beyond which the answer would be perceived as "unacceptably broad." This is an interesting issue for future research.

[3] An exception to this is the state of "exact" knowledge, in which one is able to retrieve directly a very precise answer (e.g., Answer *A* in Figure 1A) with high confidence. We assume that in such cases the answer will simply be provided "as is," with no grain adjustment needed. Therefore, this situation falls outside the scope of the current (and original) grain control model. Note that in such cases, confidence in the provided answer may be much higher than the minimum confidence criterion level, and if at ceiling (100%), confidence would no longer increase with increasing coarseness of the answer.

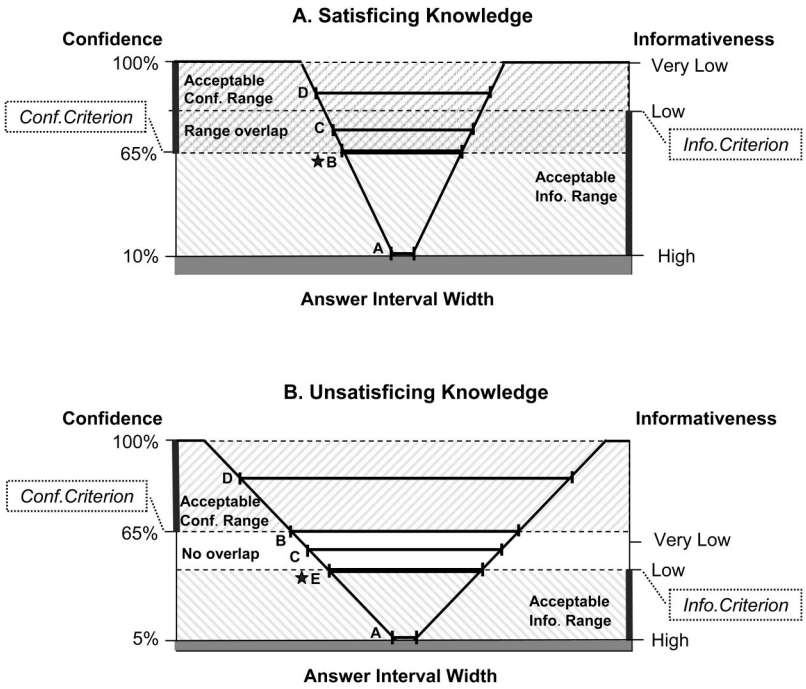

Figure 1. Schematic confidence–informativeness trade-off diagrams illustrating examples of states of satisficing knowledge (Panel A) and unsatisficing knowledge (Panel B). Candidate answers in Panel A: (A) highly informative answer with insufficient confidence, (B) the predicted chosen answer—the most precise answer that satisfies the confidence criterion and the informativeness criterion, (C) acceptable answer with respect to both criteria but not predicted by the model, and (D) coarse answer with high confidence but insufficient informativeness. Candidate answers in Panel B: (A) highly informative answer with insufficient confidence, (B) the most precise answer satisfying the confidence criterion but violating the informativeness criterion, (C) unacceptable answer with respect to both criteria (possible "compromise" answer), (D) very coarse answer with high confidence but insufficient informativeness, and (E) the predicted chosen answer—the coarsest answer satisfying the informativeness criterion but violating the confidence criterion. Conf. = confidence; Info. = informativeness; a star indicates the chosen answer according to the model.

as in Panel A. However, a much lower level of objective and subjective knowledge is represented here by a different *mapping* between the interval widths—and hence informativeness levels—of potential candidate answers, on the one hand, and confidence in the correctness of those answers, on the other. This different mapping in turn yields the defining feature of the UK state: the absence of a range of overlap between the confidence and informativeness criteria. Any answer that is coarse enough to satisfy the confidence criterion will be unreasonably coarse (e.g., Answers *B* and *D*), and vice versa, any answer that is precise enough to satisfy the informativeness criterion will be held with insufficient confidence (e.g., Answers *A* and *E*).

During the question answering process, then, respondents in a UK state will eventually reach a deadlock situation: If they begin with a precise answer (e.g., a best guess) and attempt to coarsen the answer so that it passes the confidence criterion, the process will be halted by failure to meet the informativeness criterion before reaching the confidence criterion. Alternatively, if they begin with a very broad answer (e.g., one that covers the entire range of plausible values) and attempt to make it precise enough to pass the informativeness criterion, the process will be halted by failure to meet the confidence criterion before reaching the informativeness criterion.

We propose that there are essentially two different ways of resolving this deadlock: The preferred solution might be to circumvent the criterion conflict entirely by responding "don't know" instead of providing a substantive answer. This option will be addressed in Experiment 3, below. When the don't-know option is explicitly or implicitly denied, however, at least one of the grain-selection criteria—we suggest primarily confidence—will have to be violated.

Resolving the Criterion Conflict: Confidence Criterion Violation

In some situations, the option of responding "don't know" may be unavailable, either because of social expectations or because of explicit demands to produce a substantive response (e.g., Goldsmith et al., 2002, 2005). In such situations, respondents in a UK state are in a bind: Any substantive answer must violate at least one of the two grain-selection criteria. Which criterion will respondents choose to violate—informativeness, confidence, or both?

As discussed earlier, according to the original satisficing model, which does not include an explicit informativeness criterion, the confidence criterion is maintained no matter how coarse and uninformative the answer needs to be to do so. By extension, we might assume that even though an informativeness criterion has been added to the model, once a criterion conflict arises, priority will still be given to the confidence criterion. Such an assumption would actually make the addition of the informativeness criterion to the original grain control model superfluous because the informativeness criterion would never have any effect on the grain decision, neither in the SK nor UK states. We believe, however, that respondents in a UK state will often prefer to preserve the informativeness criterion at the expense of assessed correctness. As noted by Yaniv and Foster (1997), differences in the timing of the payoffs for informativeness and correctness could induce respondents to give a higher priority to informativeness: Many social interactions are structured such that the penalty for providing an unacceptably coarse answer, and the reward for providing a precise-informative answer, is incurred immediately, whereas the fact that one has provided an incorrect answer only becomes evident at a later time, if at all. This temporal difference should encourage respondents to choose a grain size that increases their immediate gain, derived by providing a reasonably informative answer, at the expense of a possible penalty down the road if the answer turns out to be wrong.

Our hypothesis, then, is that respondents in a UK conflict state will often violate the confidence criterion in deference to the informativeness criterion. The simplest prediction is that when one is unable to satisfy both criteria simultaneously, the answer will be provided at the coarsest grain size that satisfies the informativeness criterion (see Figure 1B, Answer *E*), thereby maintaining as high a confidence level as possible, given the constraint on informativeness. Alternatively, a compromise grain size could be chosen that violates both criteria but finds a middle ground between them (see Figure 1B, Answer *C*). We will not attempt to distinguish between these two possibilities. What is important is that either of these choices, as opposed to the correctness-preserving answer (see Figure 1B, Answer *B*), constitutes a violation of the confidence criterion, which cannot be explained by the original satisficing model but is in fact predicted by the revised, dual-criterion model.

Overview of Experiments

The goal of the current study is to refine our understanding of the grain control process in answering knowledge questions under uncertainty. The preceding comparative analysis of results from Goldsmith et al. (2002, 2005) and Yaniv and Foster (1997) raised the possibility that Goldsmith et al.'s (2002) original satisficing model might be adequate under conditions in which respondents have at least moderate knowledge of the subject matter but not when respondents have only scant knowledge. We proposed that in the latter case, the inclusion of an explicit informativeness criterion, and hence the possibility of criterion conflict, is needed to explain why respondents do not always maintain high subjective correctness by providing very coarsely grained answers.

We now report a series of experiments in which we examined this proposal. In these experiments, we allowed participants full control over the grain size of their answers to factual information questions, while manipulating the level of relevant knowledge, either by comparing performance on very hard versus moderately hard sets of items (Experiments 1A, 1B, and 3) or by pre-exposing the answers to half of the very hard items (Experiment 2). We expected that a higher rate of UK-conflict states would be evoked by the low-knowledge than by the moderate-knowledge conditions. Therefore, we predicted that a higher rate of confidence-criterion violations would be observed in the former conditions than in the latter. Such a finding would embarrass the original satisficing model in favor of the dual-criterion model. In Experiment 3, the participants were given the don't-know option in addition to the control of grain size. We predicted that participants would utilize the don't-know option to circumvent the criterion conflict when in a UK state, thereby decreasing the rate of observed criterion violations, particularly in the low-knowledge condition.

Experiment 1A: Low-Knowledge Versus Moderate-Knowledge Items

In this experiment, participants answered a set of general-knowledge questions, each pertaining to some numeric-quantitative value. As in previous research (Goldsmith et al., 2002, 2005; Yaniv & Foster, 1997), this allowed grain size to be operationalized in terms of the interval widths of the answers.

The role of knowledge level in the control of grain size was examined by comparing performance on two intermixed sets of items that comprised the knowledge questionnaire. On the basis of pretesting, a set of *moderate-knowledge* (MK) items was selected, which most participants would be able to answer correctly at a "coarse, reasonably informative" grain size but not at a precise (exact value) grain size. In addition, a set of very difficult *low-knowledge* (LK) items was selected, which most participants would only be able to answer correctly at an unacceptably coarse grain size. In terms of the dual-criterion model, it was expected that the MK items would generally induce a state of SK, such that the confidence and informativeness criteria could be jointly satisfied, whereas the LK items would more often induce a state of UK, in which participants would violate the confidence criterion in favor of the informativeness criterion. In contrast, according to the original satisficing model, a common confidence criterion should be set and strictly adhered to for all items.

To examine these contrasting predictions, we used a two-phase procedure (cf. Goldsmith et al., 2002, 2005) in which participants answered each of the questions twice: In the initial, free-grain phase, the participants were allowed to answer each question by providing a precise value or whatever interval they thought would be most helpful to a "friend," given the limitations of their own knowledge. The instructions were designed to simulate a cooperative social context, in which pragmatic principles—such as Grice's (1975) maxims of quantity and quality—would apply (cf. Goldsmith et al., 2002; Yaniv & Foster, 1997). In addition, the participants were asked to provide a confidence judgment reflecting the assessed likelihood that the answer was correct. (In Experiment 1B, we verified that the elicitation of confidence judgments does not contaminate the grain control process.) In the subsequent, fixed-grain phase, the participants answered the same set of questions again (and provided confidence judgments) but this time using an interval width that was fixed in advance for each question. The interval width specified in this phase was the "coarse, reasonably informative" grain size used in selecting the MK and LK item sets. Thus, we

expected relatively high and low accuracy rates, respectively, for these two item sets.

The critical question was whether the participants would utilize the unlimited control over grain size in the initial free-grain phase to achieve the same minimum level of subjective correctness for these two sets of items, in line with the assumptions of the original satisficing model, or rather, would provide low-confidence answers when needed to achieve a minimum level of informativeness, particularly in the LK condition, in line with the dual-criterion model. Comparison of the pattern of performance for the two item types, both within and between phases, enabled us to answer this question.

Method

Participants. Twenty-four native Hebrew-speaking psychology students from the University of Haifa (Haifa, Israel) participated in the experiment for payment (NIS 35, approximately $8) or for course credit.

Materials. A 40-item general-knowledge questionnaire (in Hebrew) covering a broad range of topics was developed, in which the answer to each item was a quantitative-numeric value ("When did . . .?" "How old was . . .?" "How long is . . .?" How many . . .?" etc.). The questionnaire was comprised of 20 MK and 20 LK items. The items were selected on the basis of pretesting from an initial pool of 120 items as follows: First, in an initial pretest, the set of 120 items was presented to a panel of three independent judges (Hebrew-speaking psychology students, blind to the goals of the experiment), along with six different candidate answers whose grain sizes had been tailored for each question, ordered from precise to extremely coarse. The judges were asked to classify the candidate answers by drawing a line such that those above the line would constitute "reasonably informative" answers (of some informative value to a person asking the question), whereas those below the line would constitute "ridiculous" or "unacceptably coarse" answers (so coarse that providing such an answer would be socially prohibited). On the basis of these judgments, a *marginally reasonable* (coarse but reasonably informative) grain size was identified for each item as the coarsest grain size (of the six alternatives) that had been classified as reasonably informative by at least two of the three judges. To gauge interrater reliability, we numbered the six alternatives for each item from 1 (*narrowest answer*) to 6 (*coarsest answer*). The mean difference between the three judges in the location of the reasonable-answer cutoff line across the 120 items was 0.77 ($SD = 0.49$). This grain size was used as the "fixed" grain size in Phase 2 of the present experiment (see *Procedure* section) and also in a second pretest used to select the MK and LK items.

In the second pretest, 18 Hebrew-speaking psychology students answered the 120 questions at each of three different grain sizes: (a) a precise answer, (b) the "marginally reasonable" interval identified in the preceding pretest, and (c) an "unacceptably coarse" interval, that is, an interval that was substantially coarser than the marginally reasonable interval. The set of 20 MK items was chosen to yield a relatively high hit rate ($M = 76\%$) at the marginally reasonable grain size but a relatively low hit rate ($M = 17\%$) at the precise grain size. The set of 20 LK items was chosen to yield a relatively low hit rate ($M = 26\%$) at the marginally reasonable grain size but a relatively high hit rate ($M = 79\%$) at the unacceptably coarse grain size. These selection criteria were designed to achieve a set of MK items for which participants would generally have substantial, but not exact, knowledge, and a set of LK items for which participants would generally lack substantial knowledge but, nevertheless, have some familiarity with the subject matter so to avoid wild guessing.

Procedure. Each participant was run individually. The experiment was computer administered, except for printed instructions that were read by the participants off screen. Knowledge level was manipulated within participants by mixing the 20 MK and 20 LK items in one of four different pseudorandom orders, counterbalanced across participants. The 40 test items (and 4 additional practice items) were presented sequentially on the computer screen, in two phases.

In the first, *free-grain* phase, participants were told that they would be presented with a set of general-knowledge questions pertaining to quantitative information, and that they would be allowed to answer each question either by providing a precise number or by providing a bounded range of values of whatever width they saw fit. More specifically, they were instructed that "the principle that should guide you in your choice of answer [interval width], is that you would like your answer to be as helpful as possible to a friend who has asked you the question, while taking into account the limitations of your knowledge." Each question was then presented sequentially on screen, with two input fields appearing below the question in which participants typed in their answers: two numeric values (lower and upper bounds) if they chose to provide an interval-type answer, or a single value (typed in twice) if they decided to provide a precise answer. After confirming the answer, the entered values continued to be displayed but could no longer be changed. Participants then rated their confidence in the correctness of the answer on a scale from 0% to 100% ("What is the chance that the answer encompasses the correct value?"). Clicking the button to continue brought up the next item, and so forth, until the end of the sequence.[4]

In the second, *fixed-grain* phase, the same 40 questions (and 4 practice items) were presented again in the same sequence, but this time the participants were required to answer each question at a predefined interval width. The specified intervals were those identified on the basis of pretesting as constituting "marginally reasonable" answers (in terms of their degree of coarseness; see *Materials* section). Confidence in the correctness of these answers was also elicited, as before, on a 0%–100% scale.

Results and Discussion

Some basic results of the experiment are considered initially. For comparison, these are presented together in Figure 2. We begin with a manipulation check verifying that the LK and MK item sets indeed elicited well differentiated levels of knowledge. This was done by examining performance on the fixed-grain phase (Phase 2; Panels A and B in Figure 2), in which each question was answered at a grain size that had been specified on the basis of pretesting as representing a "coarse but reasonable" grain size for that item (see

[4] Additional exploratory data were collected by having the participants think out loud while answering the second half of the items (i.e., Items 25–44) in Phase 1. The results of these verbal protocols are not reported here. Because no performance differences were found in the results for questions that were answered out loud versus those that were answered silently, this aspect of the design is henceforth ignored.

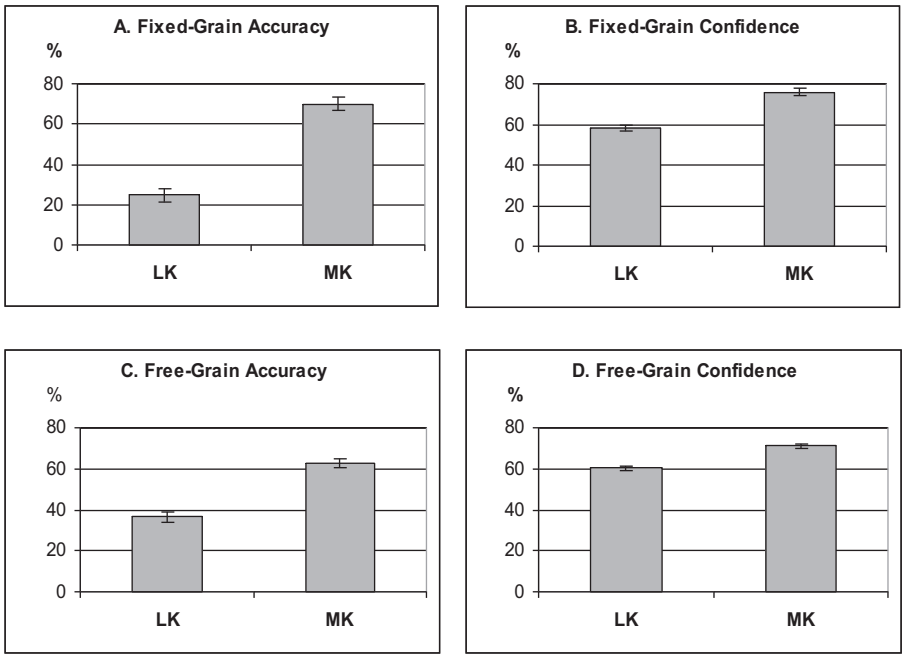

Figure 2. Mean accuracy performance (percentage correct) and subjective confidence ratings (0%–100%) for the low-knowledge (LK) and moderate-knowledge (MK) items in Experiment 1A. Error bars represent ±1 *SE* for the mean within-individual MK–LK difference.

Method section). Figure 2 (Panel A) shows that the percentage of LK items that could be answered correctly at this grain size (M = 25%) is much lower than the percentage of MK items answered correctly (M = 70%), $F(1, 23) = 198.82$, $MSE = 124.26$, $p < .001$, $\eta_p^2 = .90$ (partial eta-squared). Figure 2 (Panel B) also shows a substantial difference in comparing subjective confidence in the correctness of the LK (M = 58%) and MK (M = 76%) answers, $F(1, 23) = 137.62$, $MSE = 30.16$, $p < .001$, $\eta_p^2 = .86$. These results indicate that, indeed, the participants' level of knowledge was much lower for the LK than for the MK items, evidenced in terms of their subjective and objective ability to provide a reasonably informative correct answer for these items.

Unlike performance in the fixed-grain phase, which reflects primarily the participants' level of knowledge, performance in the free-grain phase (Phase 1; Panels C and D in Figure 2) also reflects the participants' control over the grain size of their answers. This control could potentially eliminate the differences between the MK and LK items observed in the fixed-grain phase. Nevertheless, Figure 2 (Panel C) shows that the free-grain accuracy of answers to the LK questions (M = 37%) was substantially lower than the accuracy of MK answers (M = 63%), $F(1, 23) = 110.31$, $MSE = 73.40$, $p < .001$, $\eta_p^2 = .83$. At first blush, this result may seem trivial—accuracy was lower for the more difficult, LK items than for the less difficult, MK items. However, we emphasize that when respondents have complete control over the grain size of their answers, there is no inherent reason why accuracy rates should vary with item difficulty, because participants had the option to increase the coarseness—and hence correctness—of their answers to any desired level. Indeed, according to the original satisficing model, participants are assumed to set the confidence criterion on the basis of situational incentives for correctness versus informativeness, without regard to their ability to answer a particular question. Thus, the same confidence criterion should have been applied to both LK and MK items, yielding roughly equivalent accuracy rates by providing coarser answers to LK questions than to MK questions.

How then can the original satisficing model account for such a large difference in free-grain accuracy between the LK and MK items? Perhaps part of the difference can be explained by differences in participants' ability to monitor the correctness of the LK and MK items. Although the original satisficing model assumes that respondents will control the grain size of their answers to reach a similar level of subjective correctness (i.e., confidence) for all items, whether this will translate into a similar level of actual correctness depends on the accuracy of metacognitive monitoring (e.g., Koriat & Goldsmith, 1996b). Comparison of mean subjective confidence (see Figure 2, Panel D) and actual percentage correct (see Figure 2, Panel C) indicates that overall, the participants were overconfident in the correctness of their answers, with the level of overconfidence particularly pronounced for the LK items: Overconfidence bias scores, calculated as the difference between mean confidence and actual percentage correct, averaged 23% for the LK items but only 8% for the MK items, $F(1, 23) = 41.36$, $MSE = 66.56$, $p < .001$, $\eta_p^2 = .64$ (cf. the *hard-easy effect*; e.g., Juslin, Winaman, & Olsson, 2000; Lichtenstein et al., 1982). Consequently, the difference in the subjective correctness of the MK and LK items was smaller than the actual difference that was observed. Nevertheless, a significant difference remained, with confidence in the free-grain LK answers (M = 60%) significantly lower than confidence in the free-grain MK answers (M = 71%), $F(1, 23) = 63.11$, $MSE = 22.30$, $p < .001$, $\eta_p^2 = .73$. This lower level of subjective confidence (and large effect size) indicates that not only did the free-grain LK answers

fail to reach the same level of correctness as the free-grain MK answers but the participants apparently did not intend for them to do so. The question remains, why not?

As explained earlier, the hypothesis derived from the dual-criterion model is that when the participants' knowledge is insufficient to allow them to provide an answer that is both reasonably likely to be correct (satisfying the confidence criterion) and reasonably informative (satisfying the informativeness criterion), they tend to resolve this UK conflict by violating the confidence criterion in favor of the informativeness criterion: They will provide an answer that has a relatively low assessed probability of being correct if this is needed to avoid providing an unacceptably coarse answer. Critically, such UK conflicts, and the ensuing confidence-criterion violations, should occur more often in answering LK than in answering MK items. In contrast, under the original satisficing model, confidence criterion violations, to the extent that they occur at all, should only reflect "random noise," in which case there should be no relationship between knowledge level and the rate of these violations.

To begin to evaluate this hypothesis, in Figure 3 (Panel A) we plot the categorized frequency distributions of the free-grain confidence judgments (assessed probability correct) for the MK and LK items, aggregated across participants. Clearly, the confidence distributions for the two item types differ. For example, answers with confidence ratings above 80% were more frequent for the MK than for the LK items, whereas answers with ratings lower than 80% were more frequent for the LK than for the MK items. Thus, the finding of lower mean confidence for the free-grain LK answers than for the free-grain MK answers could still be compatible with the original satisficing model, if the participants had set a common confidence criterion (e.g., 50%) that was applied with equal consistency to both LK and MK answers, but the different confidence distributions above the criterion are responsible for the different mean values.

To reject the original satisficing model in favor of the dual-criterion explanation, we must garner evidence against the idea that the confidence criterion was applied with equal consistency to the LK and MK answers. Toward this end, we began with the basic assumption of the original (and dual-criterion) model, that a common confidence criterion was set for all items, and estimated its value for each participant by two converging methods. Criterion violations were then identified as free-grain answers with subjective confidence below the estimated criterion value. We could then examine whether there was, in fact, a higher rate of criterion violations for LK than for MK items.

In the first analysis, we estimated the confidence criterion adopted by each participant on the basis of the shape of the distribution of that participant's free-grain confidence ratings. In line with the original satisfying model, we assumed that the confidence criterion used by each participant could be identified as a particular confidence level, below which there was a relatively low and stable frequency of answering (because of random noise), and immediately above which there was a sharp increase in the frequency of answering. The exact method that we used is detailed in the Appendix. The resulting criterion estimate averaged 69 (SD = 16), yielding a mean criterion violation rate (percentage of free-grain answers with below-criterion confidence ratings) of 36% (SD = 26.5). Contrary to the simple satisficing model, the violation rate was significantly higher for the LK items (M = 45%) than for the MK items (M = 27%), $F(1, 23) = 34.84$, $MSE = 102.25$, $p < .001$, $\eta_p^2 = .60$. This finding is clearly inconsistent with the assumption, on the basis of the original satisficing model, that the criterion violations were due to random noise. Instead, as predicted by the dual-criterion model, the criterion violations were particularly likely for LK items.

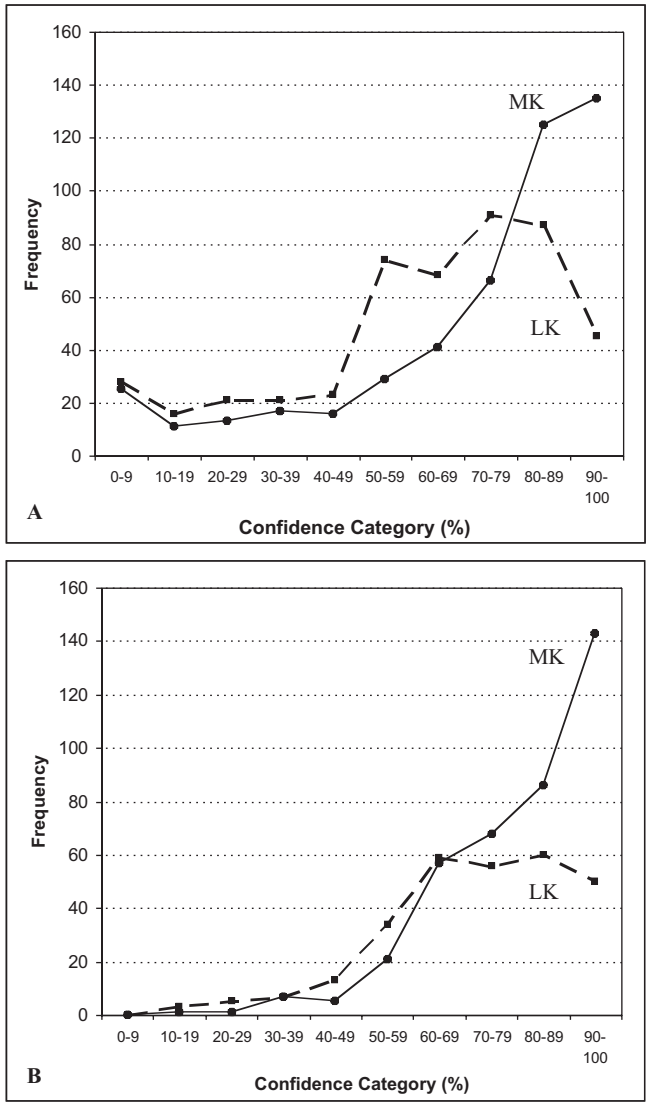

Figure 3. Categorized frequency distributions of free-grain confidence judgments (total number of answers in each category) for the low-knowledge (LK) and moderate-knowledge (MK) conditions in two experiments: (A) Experiment 1A, without report option, and (B) Experiment 3, with report option.

In the second analysis, we estimated the confidence criterion set by each participant using a theory-derived procedure adapted from the one used in previous research (Goldsmith et al., 2002, 2005; Koriat & Goldsmith, 1996b). Recall that in the second phase of the experiment, participants answered the same questions as in the first phase, but they were required to use fixed predefined intervals, identified on the basis of pretesting as coarse but still reasonably informative. By comparing the

grain size and confidence rating of each answer provided in the initial free-grain phase with the corresponding values for that item in the fixed-grain phase, it is possible to infer the location of the confidence criterion, on the basis of the assumptions of the original satisficing model. Two general categories of answers are diagnostic:

1. *The fixed-grain answer is more precise and has a lower subjective confidence rating than the free-grain answer*. This pattern implies that the subjective confidence associated with the fixed-grain answer is *below* the confidence criterion (see Figure 1A, assuming that Answer *A* is the fixed-grain answer, and Answer *B* is the free-grain answer). Otherwise, the participant would have provided the more informative, fixed-grained answer, or an even more precise answer, on the free-grain phase. That is, one would not provide a less informative, higher confidence answer than the fixed-grain answer on the free-grain phase unless confidence in the fixed-grain answer is below the criterion.

2. *The fixed-grain answer is coarser and has a higher subjective confidence rating than the free-grain answer*. This pattern implies that the subjective confidence associated with the fixed-grain answer is *above* the confidence criterion (see Figure 1A, assuming that Answer *C* is the fixed-grain answer, and Answer *B* is the free-grain answer). Otherwise, the participant would have chosen to provide the more likely-to-be-correct, fixed-grain answer, or an even coarser answer, on the free-grain phase. That is, one would not provide a more precise, lower confidence answer than the fixed-grain answer on the free-grain phase unless confidence in the fixed-grain answer is above the criterion.

These two diagnostic categories comprised 61% of the participants' answers (an average of 24 items; range = 16–35 items per participant).[5] On the basis of these items, a confidence criterion (report criterion probability-correct; P_{rc}) estimate was derived for each participant that maximized the fit between the participants' responses and the two diagnostic criteria.[6] This was done by considering all of the values between 0% and 100% as potential P_{rc} values, and then finding the P_{rc} value for each participant that maximized the P_{rc} *fit rate*, defined as the number of items in Diagnostic Category 1 for which fixed-grain confidence $> P_{rc}$ plus the number of items in Diagnostic Category 2 for which fixed-grain confidence $< P_{rc}$, divided by the total number of diagnostic items. When a range of possible P_{rc} values yielded an equivalent fit rate, the average of these values was used. This procedure yielded a mean P_{rc} estimate of 71 ($SD = 17$), with a mean fit rate of 93% (similar to previous fit rates; e.g., Goldsmith et al., 2002). Note that this mean P_{rc} estimate is very close to the mean estimate based on the shape of the confidence distribution in the previous analysis (69), and in fact the correlation between the two estimates was moderately high ($r = .62$, $N = 21$). Again we calculated the criterion-violation rate in terms of the percentage of free-grain answers with (free-grain) confidence below the estimated criterion level. Once again, contrary to the simple satisficing model, the violation rate was significantly higher for the LK items ($M = 56\%$) than for the MK items ($M = 35\%$), $F(1, 20) = 29.54$, $MSE = 145.60$, $p < .001$, $\eta_p^2 = .60$.

In sum, the converging results of these two different criterion estimation procedures are inconsistent with the assumption of the original satisficing model, that in choosing a grain size for one's answer, a common confidence criterion is set and strictly adhered to, regardless of one's level of knowledge. Instead, both the high rate of confidence-criterion violations per se, and the higher violation rate for LK than for MK items, support the predictions of the dual-criterion model, suggesting that the confidence criterion is sometimes violated to maintain a reasonable level of informativeness, and that this occurs primarily when knowledge is low.

One might wonder, however, why so many confidence criterion violations were observed for MK as well as for LK items. Apparently, the MK items also yielded a substantial number of UK-conflict states, in which participants lacked sufficient knowledge to satisfy both the confidence and the informativeness criteria. To examine this idea, we again treated the ability to answer the question correctly on the fixed-grain phase as a rough index of actual knowledge (cf. Figure 2, Panel A), and using the theory-based criterion estimates (the same pattern was obtained using the distribution-based estimates), we calculated the criterion violation rates separately for questions answered correctly and questions answered incorrectly at the fixed-grain level for each item type (LK and MK): For the LK items, there was no difference in the violation rates for items answered correctly (54%) or incorrectly (58%) in the fixed-grain phase, $F < 1$. For the MK items, however, the violation rate was substantially higher for items that were answered incorrectly on the fixed-grain phase (52%) than for those that were answered correctly (29%), $F(1, 20) = 19.53$, $MSE = 284.90$, $p < .001$, $\eta_p^2 = .49$. In fact, the violation rate for MK items for which the participants lacked knowledge (i.e., those answered incorrectly on the fixed-grain phase) was no different than for the LK items, regardless of whether these were answered correctly or incorrectly on the fixed-grain phase, $F < 1$. Presumably, the MK items that could not be answered correctly at the coarse-but-reasonable fixed-grain size yielded states of UK similar to the LK items. With regard to the MK items that were answered correctly in the fixed-grain phase, it is plausible that a certain percentage of these (e.g., the observed percentage of criterion violations) also reflect UK-states, given the marginal informativeness of the fixed-grain answers, presumed individual differences in what is considered to be a "reasonably informative" answer, and the likelihood that some of these correct coarse-grained answers are the product of guessing rather than knowledge.

Finally, one might be concerned whether there is any evidence in this experiment that the choice of grain size is based on one's confidence in the correctness of the answer, and vice versa, that the subjective confidence ratings are sensitive to the interval width of the answer that is being assessed. This issue is related to a possible alternative explanation of the observed pattern of results: If the participants' confidence ratings simply reflect, say, a gross judgment regarding overall familiarity or amount of knowledge that one has regarding the question topic, and for this reason the ratings

[5] The remaining answers were divided into three categories: 15% had equal rated confidence in both phases, 17% had increased confidence and precision on the fixed-grain phase (suggesting an increase in subjective knowledge), and 7% had decreased confidence and precision on the fixed-grain phase (suggesting a reduction in subjective knowledge).

[6] Three participants were omitted from this analysis because they had fewer than three items in one of the two diagnostic categories, making the diagnostic information unreliable. Their removal increased the intercorrelation between the theory-based and distribution-based criterion estimates but did not otherwise affect the pattern of results.

are lower for some items (e.g., the LK items) than for others on the fixed-grain phase, then for this same reason, the confidence ratings might also be lower for these items on the free-grain phase, regardless of the grain size of the answer that was chosen (for some unknown reason) by the respondent.[7] Such an account, however, is contrary to previous results showing that (a) participants tend to provide coarse grained answers when confidence in the more fine grained alternative answer is low, and (b) confidence in the correctness of these coarse-grained answers is generally higher than confidence in the correctness of the more fine-grained answers to the same questions (Goldsmith et al., 2002, 2005).

This alternative explanation is also counted against by the present results. To examine the relationship between confidence and grain size in the present experiment (see also Goldsmith et al., 2005, Experiment 2), we measured grain size as ln(interval width), a logarithmic function that approximates participants' judgments of differences in the informativeness of numeric-interval answers (Yaniv & Foster, 1995). Similar to previous findings, we found that participants tended to provide more coarsely grained answers in the free-grain phase relative to the fixed-grain phase when confidence in the correctness of the fixed-grain answer was low, and that this was true for both MK and LK items: Within-participant Pearson correlations between fixed-grain confidence and the difference in the coarseness of the free-grain answer relative to the fixed-grain answer, ln(free-grain width) − ln(fixed-grain width), averaged −.64 (SD = .16) and –.54 (SD = .25) for the MK and LK items, respectively (both means significantly different than zero, $p < .001$). In addition, a change in grain size between the two phases was associated with a corresponding change in subjective confidence: Within-participant Pearson correlations between the difference in the coarseness of the free-grain and fixed-grain answers, ln(free-grain width) – ln(fixed-grain width), and the difference in one's confidence in the correctness of the two answers, free-grain confidence – fixed-grain confidence, averaged .24 (SD = .27) and .36 (SD = .29) for the MK and LK items, respectively (both means significantly different than zero, $p < .001$). Thus, providing a more coarsely grained answer than the fixed-grain answer on the free-grain phase tended to increase confidence in the correctness of the free-grain answer relative to confidence in the fixed-grain answer, whereas providing a more precise answer than the fixed-grain answer on the free-grain phase tended to decrease confidence in the correctness of the free-grain answer.

These final analyses reinforce the theoretical assumptions common to both the original and the dual-criterion models, that the basic dynamic guiding the choice of grain size is a confidence-informativeness (or accuracy-informativeness) trade-off, leaving the original satisficing model with the problem of explaining the preponderance of low-confidence answers in the free-grain phase, and the higher rate of such answers for LK than for MK items. Our explanation, derived from the dual-criterion model and supported by the preceding results, is that these cases are confidence-criterion violations that occur when respondents are unable to provide an answer that is both reasonably likely to be correct and reasonably informative, and that they reflect the priority of the goal of informativeness over the goal of correctness.

Experiment 1B: Grain Control Without Explicit Confidence Ratings

Although confidence (subjective assessment of correctness) is postulated to be an integral part of the control of grain size in question answering, the explicit *reporting* of confidence judgments is not. Therefore, although the explicit collection of confidence judgments is a methodological necessity in the present research, it is important to determine that this aspect of the design is not responsible for the observed pattern of results. In fact, one might perhaps argue that the high rate of confidence-criterion violations in Experiment 1A was caused by the fact that the participants' answers were accompanied by an explicit confidence judgment.

To illustrate, if asked in casual conversation, "how many players are there on a hockey team?", a person might answer "I'm pretty sure it is between 4 and 12", or "I think it is between 4 and 7, but I'm not so sure." In both of these cases, the reported confidence level ("pretty sure" or "not so sure") is being conveyed as an integral part of the respondent's answer, to *qualify* the answer (Budescu & Wallsten, 1995; Smith & Clark, 1993; Teigen, 1988). The first answer would indicate that one is giving priority to correctness, whereas the second answer would indicate that one is trying hard to be informative, while warning that one is risking a wrong answer to do so. Perhaps, then, when the participants in Experiment 1A provided relatively informative free-grain answers with low confidence, from their perspective they were not actually "violating" the confidence criterion because they were treating the associated confidence judgment as a caveat. The procedure of Experiment 1A was designed to minimize this possibility by presenting the confidence rating scale to the participants only after each answer had been typed in and confirmed. Nevertheless, the question remains: Would the same rate of confidence criterion violations, and differential rate for LK versus MK items, which is the main evidence against the original satisficing model, be observed if participants provide their answers without making explicit confidence judgments?

To answer this question, we performed a control experiment. Experiment 1B was identical to Experiment 1A in all respects except one: Confidence judgments were not elicited in the free-grain answering phase. To the extent that the participants' grain choices in Experiment 1A depended on their treating the confidence judgments as an integral part of their answers, these choices should now be different in Experiment 1B. In particular, if the participants in Experiment 1A had used low confidence ratings as an explicit caveat that allowed them to provide relatively precise answers even under conditions of low (unsatisficing) knowledge, then the grain size of the answers in Experiment 1B would be expected to increase (i.e., be less informative) now that this caveat was no longer available.

Method

Participants. Twenty-four native Hebrew-speaking psychology students from the University of Haifa participated in the experiment for payment (NIS 35, approximately $8) or for course credit.

Materials. The same materials were used in Experiment 1B as in Experiment 1A.

[7] We thank Ilan Yaniv for raising this potential criticism.

Procedure. The procedure of Experiment 1B was identical to that of Experiment 1A, except that no confidence judgments were elicited in the initial, free-grain phase.

Results and Discussion

We compared the results of Experiment 1B with the corresponding results of Experiment 1A on three main performance measures: fixed-grain accuracy, free-grain accuracy, and the average grain size of the free-grain answers relative to the grain size of the corresponding fixed-grain answers, ln(free-grain width) – ln(fixed-grain width). As can be seen in Figure 4, the pattern of results for these measures in the two experiments is remarkably similar. Two-way mixed ANOVAs, Experiment × Item Type, confirmed that none of the effects or interactions involving Experiment were significant (all $Fs < 1$). The equivalence of fixed-grain accuracy between the two experiments indicates that the two groups of participants did not differ in their overall level of knowledge. More importantly, in view of the equivalent free − fixed grain size difference and equivalent accuracy of the free-grain answers in the two experiments, there is no sign that the elicitation of confidence ratings in the free-grain phase of Experiment 1A encouraged the participants to provide more precise (and hence lower confidence, less likely to be correct) free-grain answers than in the present experiment, which did not elicit confidence ratings in the free-grain phase. It seems safe to conclude, then, that the elicitation of confidence judgments in the free-grain phase was not responsible for the observed pattern of results in Experiment 1A or in the later experiments in which similar experimental procedures were used.

Experiment 2: Incidental-Learning Knowledge Manipulation

In Experiment 1A, the effect of knowledge on the grain-control process was examined by comparing the patterns of performance and subjective confidence ratings between two different sets of general-knowledge items. On the basis of the dual-criterion model, we predicted that a higher rate of confidence criterion violations would be observed for the LK compared with the MK items, because the amount of knowledge that participants could bring to bear in answering the LK questions would generally be insufficient to allow them to provide an answer that is both reasonably informative and reasonably likely to be correct. The role of knowledge, or lack of knowledge, in explaining the differential rates of confidence criterion violations is central to the dual-criterion model: According to that model, answers that violate the confidence criterion are indicative of a particular theoretical knowledge state—UK.

The purpose of Experiment 2 was to verify that the pattern of results found so far in comparing the LK and MK conditions does in fact reflect a difference in knowledge, rather than any other possible difference in the item characteristics of the LK and MK questions, such as ecological representativeness (Gigerenzer, Hoffrage, & KleinbOlting, 1991) or surface form (Cruse, 1977; Levelt & Kelter, 1982). In this experiment, a single set of very difficult general-knowledge questions was used, consisting of the LK items from Experiment 1A and similar additional items. The amount of knowledge that was available to answer these questions was manipulated by pre-exposing the participants to the precise answers to half of the questions (counterbalanced across subjects) in an initial incidental-learning phase. Incidental pre-exposure of the answers to general knowledge questions has been found to increase both the ability to answer those questions and subjective confidence in the correctness of the answers (Kelley & Lindsay, 1993). Thus, we expected this manipulation to yield a pattern of results similar to the one observed so far in comparing the MK and LK item sets: higher levels of objective and subjective knowledge for the incidentally learned items than for the unlearned items and, hence, a lower rate of criterion violations for the incidentally learned items on the free-grain phase.

Method

Participants. Twenty-four native Hebrew-speaking psychology students from the University of Haifa participated in the

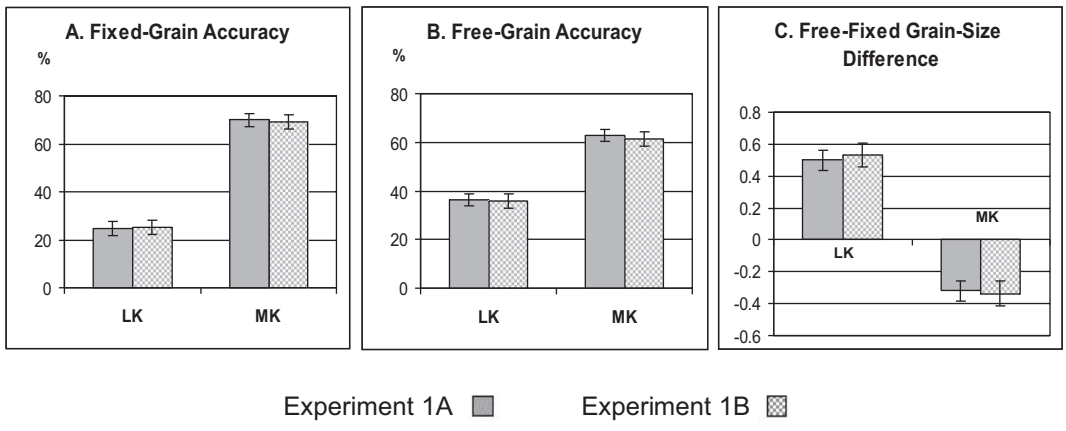

Figure 4. Mean fixed-grain accuracy, free-grain accuracy, and grain-size difference between the two phases, ln(free-grain width) − ln(fixed-grain width), in the low-knowledge (LK) and moderate-knowledge (MK) conditions of Experiments 1A and 1B. Error bars represent ±1 *SE* for the mean within-individual MK–LK difference.

experiment for payment (NIS 35, approximately $8) or for course credit.

Materials. Forty general knowledge questions were used for which a numeric answer is required (plus 4 practice questions): the 20 LK questions that were used in Experiment 1A and an additional 20 items fulfilling the same (LK) selection criteria, taken from the Experiment 1A pretest item pool (see *Materials* section of Experiment 1A).

For each of the 40 questions, an incidental-learning text segment was prepared. The segments were very short—a single sentence or paragraph of 40 words or less. Each segment presented information relating to the topic of the corresponding question, including the precise answer to the question and other related items of numeric information. Although the question itself was not presented in the text, there was enough overlap of terms and topic identifiers so as to allow easy relation between the two. Two examples follow. For the question, "How many countries are there in the African continent?" (answer: 53), the incidental-learning segment was as follows: "Africa includes 53 countries. Most of these were established during the second half of the 20th century, for example, Lesotho and Benin, which were established during the '60s." For the question, "What was the population of the city of Haifa in the year 2000" (answer: 270,500), the incidental-learning segment was as follows: "Haifa's population increased from about 250,000 in 1990 to 270,500 in 2000." As in these examples, many of the segments included coarse numeric approximatives (e.g., "about 250,000," "during the '60s") in addition to precise numeric values, to signal to the participants that some vagueness is acceptable in the current context and to resemble informal communication rather than an encyclopedic listing of information.

Procedure. The experimental procedure was the same as in Experiment 1A, except that an incidental-learning phase preceded the other two phases. In this phase, 20 text segments were presented, and the participants were asked to rate the novelty of the information contained in each segment, thereby incidentally pre-exposing them to the answers to half of the subsequent knowledge-test questions. Items whose answers were pre-exposed for one half of the participants were not pre-exposed for the other half of the participants, and vice versa, so that all items were used equally often in the incidental-learning and no-incidental-learning conditions. Participants were asked to read each segment carefully and to rate the novelty of the information conveyed in the segment using one of four options: "completely new," "mostly new," "somewhat new," or "not new." After this phase, the remainder of the procedure was the same as in Experiment 1A: two test phases, a free–grain phase followed by a fixed-grain phase, both including the collection of confidence ratings. The same order of items was used in each phase, with the pre-exposed and non-pre-exposed items randomly intermixed.

Results and Discussion

The incidental learning manipulation in this experiment was intended to allow a comparison of performance under conditions of moderate versus low knowledge, which parallels the comparison made in Experiment 1A using the two different item sets (MK vs. LK). Hence, for the sake of continuity with our previous terminology, we refer to items whose answers were pre-exposed as *MK items* or the *MK condition*, and to items whose answers were not pre-exposed as *LK items* or the *LK condition*, while reminding the reader that in this experiment the same items served equally often in each condition.

Some basic results of the experiment, to be considered initially, are presented in Figure 5. As can be seen by inspecting the figure (Panels A and B), the incidental-learning manipulation was successful in creating two different levels of objective and subjective knowledge: The accuracy of the fixed-grain answers to the pre-exposed, MK items ($M = 56\%$) was substantially higher than the accuracy of the answers to the non-pre-exposed, LK items ($M = 22\%$), $F(1, 23) = 101.17$, $MSE = 132.49$, $p < .001$, $\eta_p^2 = .82$. Accordingly, subjective confidence in the correctness of the fixed-grain MK answers ($M = 71\%$) was higher than confidence in the correctness of the fixed-grain LK answers ($M = 55\%$), $F(1, 23) = 40.15$, $MSE = 71.75$, $p < .001$, $\eta_p^2 = .64$. Although some of this 16 percentage-point confidence difference may be the result of increased familiarity of the pre-exposed questions (e.g., Koriat & Levy-Sadot, 2001), in light of the much larger, 34 percentage-point difference in actual accuracy of the answers, it would appear that the incidental learning manipulation was primarily affecting objective knowledge, which in turn was affecting subjective knowledge.

The original satisficing model holds that by exercising control over the grain size of their answers in the free-grain phase, participants should, in principle, be able to eliminate the accuracy and confidence differences between the LK and MK conditions observed in the fixed-grain phase. Nevertheless, as in Experiment 1A, the accuracy of the free-grain LK answers (38%; see Figure 5, Panel C) was substantially lower than the accuracy of the free-grain MK answers (57%), $F(1, 23) = 40.43$, $MSE = 100.94$, $p < .001$, $\eta_p^2 = .64$. Moreover, confidence in the correctness of the free-grain LK answers ($M = 60\%$) was significantly lower than confidence in the correctness of the free-grain MK answers ($M = 75\%$), $F(1, 23) = 37.00$, $MSE = 77.60$, $p < .001$, $\eta_p^2 = .62$ (see Figure 5, Panels C and D). This again suggests, as in Experiment 1A, that the participants did not intend to reach the same level of accuracy in answering the LK and MK questions.

As explained in analyzing the results of Experiment 1A, to reject the original satisficing model in favor of the dual-criterion explanation, we must show that the difference in mean confidence between free-grain LK and MK answers does not derive merely from differences in the distributions of the confidence ratings for LK and MK answers that passed some common confidence criterion. We shall again do so by showing that the confidence criterion was not applied equally to all items, and in particular, that participants chose to violate the confidence criterion more often in response to LK questions than in response to MK questions.

The same two procedures used in Experiment 1A to estimate the confidence criterion set by each participant, and the ensuing criterion violation rates, were used again here. In the first analysis, we estimated the confidence criterion adopted by each participant on the basis of the shape of the distribution of that participant's free-grain confidence ratings (see the Appendix). The resulting criterion estimates averaged 68 ($SD = 15$), yielding a mean criterion violation rate (percentage of free-grain answers with below-criterion confidence ratings) of 35% ($SD = 27$). Importantly, contrary to the simple satisficing model, the violation rate was significantly higher for the LK items ($M =$

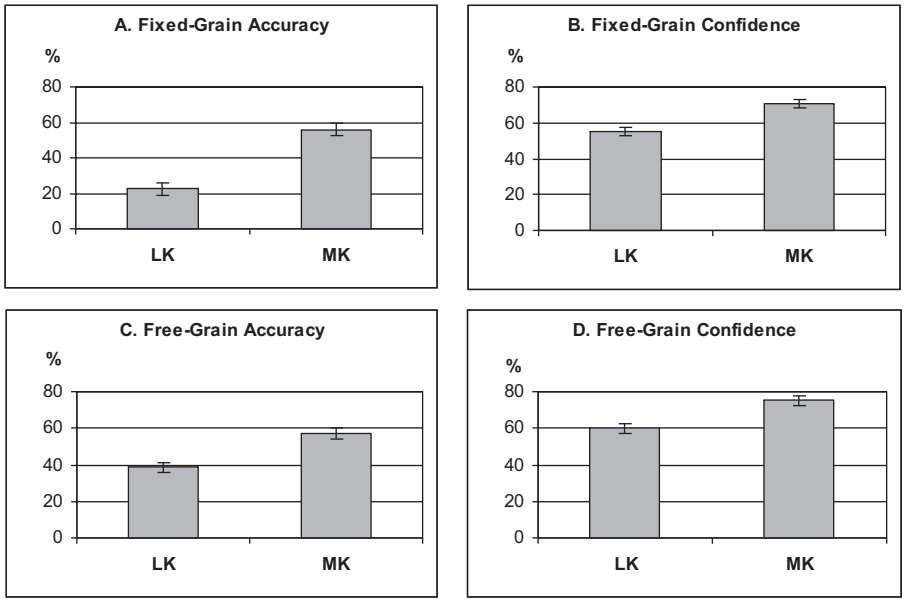

Figure 5. Mean accuracy performance (percentage correct) and subjective confidence ratings (0%–100%) for the non-pre-exposed (LK = low-knowledge) and pre-exposed (MK = moderate-knowledge) items in Experiment 2. Error bars represent ±1 *SE* for the mean within-individual MK–LK difference.

45%) than for the MK items (M = 25%), $F(1, 23) = 27.68$, $MSE = 178.6$, $p < .001$, $\eta_p^2 = .55$.

In the second analysis, we inferred the location of the confidence criterion adopted by each participant on the basis of a set of theoretically diagnostic items: (1) items for which the fixed-grain answer is more precise and has a lower subjective confidence rating than the free-grain answer (in which case fixed-grain confidence should be below the criterion), or (2) items for which the fixed-grain answer is coarser and has a higher subjective confidence rating than the free-grain answer (in which case fixed-grain confidence should be above the criterion). In this experiment, 52% of the participants' answers (an average of 21 items per participant; range = 11–32 items) fell into one of these two diagnostic categories.[8] The estimated confidence criterion (P_{rc}) that maximized the fit between each participant's responses and the diagnostic criteria averaged 77 (SD = 13), with a mean fit rate of 91%. This mean "theory-based" P_{rc} estimate is somewhat higher than the estimate based on the shape of the confidence distribution in the previous analysis (68). Nevertheless, as in Experiment 1A, good convergent validity is indicated by a moderately high correlation between the two estimates ($r = .60$, $N = 20$). By this estimate too, contrary to the simple satisficing model, the confidence criterion violation rate was significantly higher for the LK items (M = 76%) than for the MK items (M = 48%), $F(1, 19) = 32.35$, $MSE = 46.38$, $p < .001$, $\eta_p^2 = .63$.

Again, converging results based on the two criterion estimation procedures lead us to reject the assumption of the original satisficing model, that in choosing a grain size for one's answer, a common confidence criterion is set and strictly adhered to regardless of one's level of knowledge. Because the violation rate in the MK condition of this experiment was even higher than in Experiment 1A, we again examined the possibility that these violations also reflect states of UK, despite the incidental learning manipulation. Recall that in this experiment, the MK items were actually the LK items from Experiment 1A (or similar, very difficult questions) whose answers were incidentally pre-exposed. Calculating the criterion violation rates separately for items answered correctly and those answered incorrectly in the fixed-grain phase (on the basis of the theory-based criterion estimates; distribution-based estimates yielding a similar pattern), we found that the violation rate for MK items answered incorrectly on the fixed-grain phase (69%) was substantially higher than when the items were answered correctly on the fixed-grain phase (33%), $F(1, 19) = 61.81$, $MSE = 209.41$, $p < .001$, $\eta_p^2 = .76$. A smaller difference was also found in the violation rates of LK items answered incorrectly (79%) versus correctly (64%) in the fixed-grain phase, $F(1, 19) = 8.31$, $MSE = 258.54$, $p < .01$, $\eta_p^2 = .30$. Thus, whether the question could be answered correctly in the fixed-grain phase accounted for a very substantial amount of the variance (as reflected by η_p^2) in the violation rate for MK items (76%), and a smaller amount for LK items (30%). As in Experiment 1, we presume that much of the residual violation rate for correct fixed-grain MK items (33%, similar to Experiment 1A) can be explained in terms of guessing (low subjective knowledge), with confidence in correct fixed-grain items that violated the estimated confidence criterion averaging 69%, compared with 88% for correct fixed-grain items that did not violate the confidence criterion.

[8] As in Experiment 1A, 4 participants were omitted from this analysis because they had fewer than three items in one of the two diagnostic categories. Their removal did not affect the pattern of results.

Experiment 3: Using "Don't Know" to Circumvent the Criterion Conflict

The experiments reported so far have examined the control of grain size in question answering in a situation in which the respondent has complete control over the grain size of his or her answers but is required to provide a substantive answer to each question. In real-life situations, however, respondents generally have available a further means of control, which has been called *report option*—the option to respond "don't know" or "don't remember" rather than provide a substantive answer (Goldsmith & Koriat, 1999, 2008; Koriat & Goldsmith, 1994, 1996b). Previous research has shown that respondents, even young children, can utilize report option to enhance the accuracy of the information that they report by withholding answers about which they are unsure (e.g., Higham, 2007; Kelley & Sahakyan, 2003; Koriat & Goldsmith, 1994, 1996b; Koriat, Goldsmith, Schneider, & Nakash-Dura, 2001; Rhodes & Kelley, 2005; Roebers & Schneider, 2005). In those studies, control over the grain size of the answers was either prevented or minimized.

What should we expect when respondents are given full control over the grain size of their answers and the option to withhold the answer entirely? Under such conditions, one might wonder whether the participants would ever utilize the option to withhold the answer. After all, even if they had little or no knowledge regarding the question, they might still prefer to provide a very coarse answer that conveys some information rather than withhold the answer and provide no information at all. This, in fact, is the prediction that might be derived from the original satisficing model. According to that model, when asked, for example, "How many countries are there in the African continent?", if the respondent needs to provide a very coarse answer—such as "between 5 and 200"—to pass the confidence criterion, that is what he or she will do.

In contrast, according to the dual-criterion model, respondents should provide a substantive answer when they are able to simultaneously satisfy both the minimum confidence criterion and the minimum informativeness criterion, but they should refrain from answering otherwise. The assumption is that in comparison with providing an extremely uninformative answer that violates the informativeness criterion, such as "between 5 and 200" in the previous example, the admission of ignorance is a more normative, socially acceptable response. We assume that responding "don't know" is also more normative than providing an informative answer that has an unreasonably low subjective probability of being correct (i.e., one that violates the confidence criterion). Thus, because it offers a way to circumvent the UK conflict entirely, providing a don't-know response should generally be preferable to violating one or both of the report criteria.

To decide between the two models, and in particular, to examine the postulated role of report option in circumventing criterion conflicts under the dual-criterion model, in Experiment 3 we used the same procedure as in Experiment 1A (with the same MK and LK item sets), except for the addition of the option to respond "don't know" in the initial, free-grain phase. In that phase, participants could provide a precise answer, an interval answer (of any width), or withhold the answer entirely. As just explained, according to the original satisficing model, given that participants have full control over the grain size of their answers, they actually have no need for the don't-know response, and hence, if used at all, the choice of this response should be entirely arbitrary. In contrast, our hypothesis, derived from the dual-criterion model, was that the participants would utilize the don't-know option systematically to resolve UK-conflict states. Hence, don't-know responses should be observed more often in the LK condition than in the MK condition, and primarily when subjective confidence in the fixed-grain answer is low. Moreover, as a result of the opportunity to avoid criterion violations by using the don't-know response, there should be a substantially lower rate of confidence criterion violations in this experiment than in Experiment 1A (in which the don't-know option was unavailable) and a smaller difference in the violation rates between the MK and LK conditions. Of course, if the don't-know response was used to resolve *all* UK-conflict states, no criterion violations should be observed at all in this experiment. However, there may also be a social-normative prohibition against invoking the don't-know option too often, which might constrain its use.

Method

Participants. Twenty-four native Hebrew-speaking psychology students from the University of Haifa participated in the experiment for payment (NIS 35, approximately $8) or for course credit.

Materials. The same 40 questions (20 MK and 20 LK) used in Experiment 1A were used in this experiment.

Procedure. The experimental procedure was the same as for Experiment 1A, comprised of two phases, free-grain and fixed-grain. The only change was in the initial, free-grain phase, in which participants were now allowed to respond "don't know": Participants could type in a precise or interval-type answer, if they felt that was the most appropriate response, or they could click on a separate "don't know" response option. The same instructions used to guide the participants' answering in Experiment 1A (being helpful to a friend given the limitations of one's own knowledge) were used again here. No confidence ratings were collected for don't-know responses in the free-grain phase.

Results and Discussion

To allow comparison of the grain-control process in this experiment, which included the don't-know option, with the grain-control process observed in Experiment 1A, which did not include this option, we first compared performance in the two experiments on the fixed-grain phase to ensure that the two groups of participants had similar levels of objective and subjective knowledge. With regard to objective knowledge, fixed-grain accuracy was virtually identical in the two experiments, both for the LK items (25% in both experiments) and for the MK items (67% in Experiment 3 vs. 70% in Experiment 1A, $F < 1$). As in Experiment 1A, the difference in fixed-grain accuracy between the two item types was substantial, $F(1, 23) = 134.90$, $MSE = 151.36$, $p < .001$, $\eta_p^2 = .85$. Subjective knowledge, as expressed by fixed-grain confidence, was also similar in the two experiments, both for the LK items (53% in Experiment 3 vs. 58% in Experiment 1A), $F(1, 46) = 1.23$, $MSE = 192.82$, *ns*, and for MK items (69% in Experiment 3 vs. 76% in Experiment 1A), $F(1, 46) = 4.01$, $MSE = 145.69$, *ns*. Note that although the difference was not statistically

significant, if anything, the slightly lower levels of subjective knowledge observed in this experiment relative to Experiment 1A would tend to increase the number of UK-conflict states and, hence, to increase the rate of confidence-criterion violations (contrary to the predictions of the dual-criterion model, and contrary to the actual results reported below). As with the objective knowledge measure, the difference in fixed-grain confidence between the two item types in this experiment was substantial, $F(1, 23) = 52.72$, $MSE = 58.76$, $p < .001$, $\eta_p^2 = .70$.

Turning now to the free-grain phase, in which participants had full control over the grain size of their answers, as well as the option to refrain from answering entirely (i.e., to respond "don't know"), the prediction derived from the original satisficing model, that participants would have no need for the don't-know option, was refuted. Although they could, in principle, have provided a very (unacceptably) coarse answer with a very high probability of being correct (e.g., "between 0 and 10 million"), the participants chose instead to give don't-know responses to over one third of the questions, and, as predicted by the dual-criterion model, they did so more often to the LK questions ($M = 45\%$) than to the MK questions ($M = 30\%$), $F(1, 23) = 18.81$, $MSE = 5.27$, $p < .001$, $\eta_p^2 = .45$. According to the dual-criterion model, participants should invoke the don't-know option for those cases in which they are unable to provide a reasonably informative answer with a reasonably high level of subjective confidence (i.e., in UK-conflict states). Consistent with this idea, the within-participant Goodman–Kruskal gamma correlation between subjective confidence in the fixed-grain answer and whether it would be withheld on the free-grain phase averaged –.64 and –.77 for the LK and MK item sets, respectively, and –.71 across all items.

Just as the availability of grain control did not obviate the use of report option, the availability of report option did not obviate the control of grain size. Thus, among the answers that the participants chose to provide on the free-grain phase, there was also a negative relationship between subjective confidence in the fixed-grain answer and the relative coarseness of the free-grain answer (indexed, as before, in terms of ln[free-grain width] – ln[fixed-grain width]), with the within-participant Pearson correlation between these two variables averaging –.54 ($SD = .27$) and –.61 ($SD = .27$) for the LK and MK item sets, respectively, and –.64 ($SD = .18$) across all items. In fact, in comparing the control of grain size between the two knowledge conditions (LK and MK), in the LK condition the free-grain answers tended to be coarser than the fixed-grain answers (relative coarseness averaging .43), whereas in the MK condition the free-grain answers tended to be more precise than the fixed-grain answers (relative coarseness averaging –.26), $F(1, 23) = 45.32$, $MSE = 0.13$, $p < .001$, $\eta_p^2 = .66$, for the difference between conditions.

Given the opportunity to utilize both report option and grain control differentially between the two knowledge conditions, one might have expected that in this experiment, unlike in the previous experiments, the participants would achieve equivalent free-grain accuracy for the answers to LK and MK items. This was not the case, however: The free-grain LK answers ($M = 39\%$) were substantially less accurate than the free-grain MK answers ($M = 65\%$), $F(1, 23) = 35.92$, $MSE = 223.43$, $p < .001$, $\eta_p^2 = .61$. The difference in the subjective correctness of the free-grain LK answers ($M = 69\%$) and MK answers ($M = 79\%$) was smaller but no less significant, $F(1, 23) = 49.44$, $MSE = 22.96$, $p < .001$, $\eta_p^2 = .68$. As in the previous experiments, the question remains, then, whether this difference stems from a higher rate of confidence-criterion violations for LK items compared with MK items, or from potentially different distributions of confidence in the correctness of answers that are all above the confidence criterion.

The aggregate distribution of the confidence ratings in Experiment 3 (free-grain answers remaining after the exercise of report option) is plotted in Figure 3 (Panel B, earlier), for comparison with the corresponding plot for Experiment 1A (free-grain answers without report option; see Figure 3, Panel A). Apparently because report option was used to withhold low-confidence answers, particularly in the LK condition, the frequency of confidence levels lower than 40% was attenuated to very few cases, and the large difference between the item sets in the range of 40%–80% confidence observed in Experiment 1A was reduced as well. Now the main difference between the frequency of LK and MK answers appears to be in the high range of confidence ratings, between 70 and 100, which are likely to lie above the confidence criterion.

We again estimated the confidence criterion adopted by each participant, and the ensuing violation rates, using the two converging methods. As before, the first analysis estimated the confidence criterion adopted by each participant on the basis of the shape of the distribution of that participant's free-grain confidence ratings (see the Appendix). The resulting criterion estimates averaged 72 ($SD = 16$), which is equivalent to the mean estimate in Experiment 1A (69), $F < 1$. Nonetheless, in a two-way analysis, Experiment × Item Type, the violation rate in this experiment (17% of the questions) was much lower than in Experiment 1A (35%), $F(1, 46) = 10.14$, $MSE = 791.1$, $p < .01$, $\eta_p^2 = .18$. Moreover, there was a significant interaction between experiment and knowledge condition, such that the violation rate was reduced more for LK items than for MK items, $F(1, 46) = 10.37$, $MSE = 88.5$, $p < .01$, $\eta_p^2 = .18$. Hence, compared with Experiment 1A, in which a much higher rate of criterion violations was observed for LK (44%) than for MK (27%) items, in this experiment there was only a small difference between the two item types (20% vs. 15%, respectively), which did not reach statistical significance, $F(1, 23) = 4.19$, $MSE = 77.7$, $p < .06$.

A similar pattern was found using the theory-derived criterion estimates, on the basis of diagnostic items in which the fixed-grain answer was more precise and held with lower confidence than the free-grain answer, or vice versa. Because many of the free-grain answers were withheld in this experiment, there were fewer diagnostic items per participant than in the previous two experiments (61% of the answered items; $M = 15$ diagnostic items; range = 7–25). Nevertheless, once again the mean theory-based criterion estimate (77, $SD = 17$, mean *fit rate* = 97%) was close to the mean distribution-based estimate, with convergent validity indicated by a moderately high correlation between the two estimates ($r = .67$, $N = 24$).[9]

Again in a two-way analysis, Experiment × Item Type, the overall theory-based violation rate in this experiment (32% of the questions) was significantly lower than in Experiment 1A (52%),

[9] One participant had no items in one of the two diagnostic categories, accompanied by a very high criterion estimate. Because the distribution-based estimate for this participant was equally high, her data were retained in the analysis.

$F(1, 43) = 11.33$, $MSE = 807.3$, $p < .01$, $\eta_p^2 = .21$. No less important is the significant interaction between experiment and knowledge condition, indicating that the violation rate was reduced more for LK items than for MK items, $F(1, 43) = 11.50$, $MSE = 145.5$, $p < .01$, $\eta_p^2 = .21$. Consequently, whereas in Experiment 1A a substantially higher rate of criterion violations was observed for LK (63%) than for MK (41%) items, in this experiment the two violation rates were equivalent (34% and 29%, respectively), $F(1, 23) = 1.74$, $MSE = 158.1$, *ns*. Also, unlike Experiment 1A (and Experiment 2), in this experiment there was no difference in violation rates for MK items that could be answered correctly (29% violations) versus incorrectly (30% violations) in the fixed-grain phase ($F < 1$).

In sum, the converging patterns of results from both the theory-based and distribution-based analyses suggest that, as predicted by the dual-criterion model, the participants utilized the don't-know response to resolve criterion conflicts primarily when their level of knowledge was low. This reduced the frequency of free-grain answers provided with low (subcriterion) confidence but did not eliminate them entirely, with residual confidence-criterion violation rates of 17% by the distribution-based analysis and 32% by the theory-based analysis. Some of these may be due to measurement error (i.e., inaccurate estimation of the actual criterion used by some participants). We speculate, however, that some of the remaining criterion violations, despite the option to withhold them entirely, are real, perhaps reflecting a reluctance to overuse the don't-know option. That is, it may be socially acceptable to admit ignorance, as long as one does not do so too often.

General Discussion

The present research advances a line of work on the strategic regulation of memory accuracy and informativeness (Goldsmith & Koriat, 2008). So far, the theoretical models that have evolved from this work have emphasized correctness (accuracy) more than informativeness (quantity): Both in Koriat and Goldsmith's (1996b) model of the control of report option, and in Goldsmith et al.'s (2002) satisficing model of the control of grain size, the goal of correctness is handled explicitly, by setting and satisfying a minimum-confidence criterion, whereas the goal of informativeness is strived for implicitly, either by volunteering all answers that pass the criterion (in the case of report option) or by providing the most precise answer that passes the confidence criterion (in the case of grain size). With respect to the control over grain size, the present research calls into question this basic asymmetry. As put forward in the newly proposed dual-criterion model, in addition to protecting the accuracy objective with an explicit minimum-confidence criterion, it appears that rememberers also protect the informativeness objective with an explicit minimum-informativeness criterion. The full role of the informativeness objective, which remained hidden in previous studies, was revealed in this research by creating a situation in which respondents have low levels of knowledge.

The inadequacy of a grain-selection process that is controlled solely by a confidence criterion, as embodied in the original satisficing model, was indicated by the basic finding of Experiments 1A and 2: A substantial percentage of answers was provided in the free-grain phase with low levels of confidence that violated the estimated confidence criterion for each participant. In principle, the confidence criterion could have been passed simply by providing coarser answers to these questions. Yet, the participants chose not to do so, apparently because such answers would be so uninformative as to be prohibited by social-pragmatic norms of communication, in violation of the minimum-informativeness criterion. Thus, the second important finding of these experiments was that the observed confidence-criterion violations were not simply the result of random noise but, rather, were systematic: As predicted by the dual-criterion model, the rate of these violations was inversely related to the respondent's level of knowledge regarding the question, with a higher violation rate observed when comparing performance on very difficult versus moderately difficult sets of questions (Experiment 1A), and when knowledge level was manipulated by pre-exposing the answers to half of the items (Experiment 2). This relationship was also observed within the MK conditions when comparing questions answered correctly versus those answered incorrectly on the second, fixed-grain phase (Experiments 1A and 2). The ability to answer an item correctly in the "coarse but still informative" fixed-grain phase appeared to provide a more sensitive measure of whether each particular participant had or lacked SK for each particular question. Overall, these results support the dual-criterion model and its claim that confidence-criterion violations should occur specifically when one is unable to provide an answer that is both sufficiently informative and sufficiently likely to be correct.

Further support for the existence of a minimum-informativeness criterion and its role in constraining the grain-control process in cases of low knowledge was obtained in Experiment 3. In that experiment, participants were given the option to respond "don't know," in addition to having full control over the grain size of their answers. According to Koriat and Goldsmith's (1996b) report-option model, which did not include control over grain size, respondents exercise the don't-know option when confidence in the correctness of their candidate answer falls below the report criterion. Given full control over grain size, however, respondents could always coarsen their answers enough to pass the criterion, no matter how high it might be. Thus, there would be no reason for respondents ever to utilize the don't-know option when they are given full control over grain size as well. Yet, the participants in Experiment 3 did invoke the don't-know option, in responding to about 38% of the questions, indicating that the control of report option still served a purpose. As discussed next, with its inclusion of a minimum-informativeness criterion, the dual-criterion model is able to clarify that purpose.

Toward a Unified Model of Grain Size and Report Option

Most of the research carried out so far on the control of grain size and report option in conveying information from memory has allowed the participants only one type of control (grain size *or* report option) while denying them the other. Although this strategy has clear advantages in terms of increased experimental control, it comes at the price of reduced ecological validity. When answering questions under uncertainty in real-life contexts, people generally have both the option to control the coarseness of their answer, and the option of withholding it

entirely, if that seems most appropriate. How do they manage the utilization of both types of control simultaneously, deciding between the various alternatives?

On the basis of the dual-criterion model of grain control, we hypothesized that the don't-know option would be invoked, despite full control over grain size, whenever the coarsening of one's answer to satisfy the confidence criterion makes the answer unacceptably uninformative, that is, the answer becomes so coarse that it violates the minimum-informativeness criterion. Because one's knowledge is insufficient to allow the confidence criterion to be met without violating the informativeness criterion, with control over grain size alone, the respondent is in a bind. With the additional control of report option, however, the don't-know response can be used to circumvent the deadlock. On this view, a "don't-know" response has a specific metacognitive meaning. It does not mean that one cannot convey any information at all in response to the question. Rather, it means that one's knowledge is insufficient to provide an answer that is both reasonably likely to be correct and reasonably informative in the current communication context; it signals that one is in a state of UK.

Consistent with this idea, the frequency of observed confidence-criterion violations in Experiment 3, in which the don't-know option was available, was substantially lower than in Experiment 1, in which this option was denied. This reduction was particularly pronounced for the LK items and for items that the participants were unable to answer correctly in the fixed-grain phase (i.e., at a coarse but reasonably informative grain size). Thus, the don't-know response tended to be invoked selectively when knowledge was poor, in cases that otherwise would have necessitated the respondent to violate one or both of the report criteria.

A somewhat surprising finding, however, was the significant rate of confidence-criterion violations that remained in Experiment 3 (17% of all items by the distribution-based analysis and 32% by the theory-based analysis) despite the fact that participants could have chosen to respond "don't know" to these items instead of providing substantive answers. As mentioned earlier, the absolute levels of these rates may be inflated somewhat by measurement error (i.e., inaccurate estimation of the actual criterion used by some participants), particularly in the theory-based analysis, which was based on relatively few items per participant after the don't-know responses were discarded. We suspect, however, that a significant portion of these criterion violations are real, reflecting a reluctance to overuse the don't-know option, which also stems from social-pragmatic considerations related to informativeness but at the more global level of the communication episode as a whole.

To clarify this idea, consider again the implication of Koriat and Goldsmith's (1996b) report-option model, mentioned earlier, that if one's knowledge regarding a set of questions is so poor that one cannot produce any answer with enough confidence to pass the confidence criterion, then one will simply not volunteer any answers, responding "don't know" to all of the questions. Surely someone in such a situation will feel uncomfortable repeatedly responding "don't know," because of the implicit expectation that one should have at least some knowledge of the topic and, hence, that one should be able to convey at least some minimal amount of information. Thus, in addition to the minimum-informativeness criterion for individual answers considered so far, there may also be more global informativeness considerations that operate across sets of questions or communication episodes: Respondents are expected to provide substantive answers to at least some of the questions that they are asked. Yet, in situations of low knowledge, doing so may require them to violate the confidence (or informativeness) report criterion at the level of individual answers. Although speculative, this idea could perhaps explain situations in which people who remember little, such as elderly people or people being tested after a long delay, adopt a more liberal report criterion than people who remember more (Kelley & Sahakyan, 2003; Koriat & Goldsmith, n.d. [unpublished data]; Pansky, Goldsmith, Koriat, & Pearlman-Avnion, in press).

In sum, although there is still a need for more work, the present study represents a significant step toward an integrated model of report option and grain size. The findings indicate that any such model will need to include an explicit minimum-informativeness criterion as well as a minimum-confidence criterion, and that the (in)ability to jointly satisfy both criteria may be a key factor that determines which type of control is utilized. The results also suggest the need to consider subjective goals that go beyond the accuracy and informativeness of individual answers. For example, as just discussed, the finding that a significant number of confidence criterion violations remained in Experiment 3 even though participants could have responded "don't know" to these questions, may perhaps reflect the influence of personal-social expectations for responsiveness at a more global, "conversational" level.

Toward an Expanded View of Metacognitive and Social Processes in Remembering

A final, more general implication of the present work is the need to incorporate a wider range of phenomena and processes into the study of memory than has been done in traditional memory research. A great deal of research on metacognition over the past decades has examined the processes by which people monitor the validity of their memories, as well as the accuracy of this monitoring (e.g., Gigerenzer et al., 1991; Koriat, 1993; Koriat, Lichtenstein, & Fischhoff, 1980; Schwartz, 1994). More recently, attention has turned to the role of metacognitive processes in remembering, and in regulating memory performance (e.g., Barnes et al., 1999; Dodson & Schacter, 2002; Goldsmith & Koriat, 1999, 2008; Higham, 2002, 2007; Koriat & Goldsmith, 1994, 1996b; Koriat et al., 2008; Mitchell & Johnson, 2000; Nelson & Narens, 1990, 1994; Odegard & Lampinen, 2006; Perfect & Schwartz, 2002; Reder, 1987, 1988). Thus, for example, the examination of monitoring and control processes operating during memory retrieval has provided important insights regarding developmental changes in memory accuracy (e.g., Koriat et al., 2001; Lindsay, Johnson, & Kwon, 1991; Roebers & Schneider, 2005), memory deficits in old age (e.g., Henkel, Johnson, & De Leonardis, 1998; Jacoby, Bishara, Hessels, & Toth, 2005; Kelley & Sahakyan, 2003; Koutstaal, 2006; Rhodes & Kelley, 2005), memory impairment in clinical populations (Danion, Gokalsing, Robert, Massin-Krauss, & Bacon, 2001; Gilboa et al., 2006; Koren, Sneidman, Goldsmith, & Harvey, 2006), and strategic control of psychometric exam performance (Higham, 2007; Higham & Arnold, 2007; Koriat & Goldsmith, 1998).

Beyond such extensions, however, the present work implies that additional types of monitoring processes, and their contributions to memory performance, must be addressed as well (see also Jost, Kruglanski, & Nelson, 1998), in particular, the role of perceived informativeness. Bringing this factor into the spotlight, a new component is added to metacognitive memory theory, leading to further challenging questions: How do rememberers evaluate the informativeness of potential candidate answers? On what basis do they set the minimum-informativeness criterion? What are the factors that determine how criterion conflicts are resolved? Some of the complexities involved in these questions can be illustrated by examples from the pretest data used in choosing the general-knowledge items for our experiments: With regard to the question, "When was Franklin Roosevelt first elected president of the United States?" (correct answer: 1933), "1920–1950" was judged to be a reasonably informative answer, whereas "1900–1980" was judged to be unacceptably coarse. At the same time, with regard to the question, "When did David Ben-Gurion immigrate to (what is now) Israel?" (correct answer: 1906), "1900–1910" was judged to be a reasonably informative answer, whereas "1900–1930" was perceived as unacceptably coarse. Notice, then, that a grain size (interval width) of 30 years was judged by Israeli participants as being reasonably informative with regard to the election date of a U.S. president but unacceptably coarse with regard to the immigration date of Israel's first prime minister, a fact studied by every Israeli high school student. Presumably, the results for these two questions would have been quite different (e.g., reversed) had the judgments been made by U.S. participants. These and many other examples (e.g., judging a grain size of 15 to be reasonably informative regarding the number of African countries but unacceptably coarse regarding John F. Kennedy's age at the time of his assassination) imply that the subjective evaluation of informativeness is both content and context specific. It should also be sensitive to the perceived level of knowledge of the recipient of the information (Grice, 1975; Sperber & Wilson, 1995): An Israeli student might feel comfortable providing a 30-year estimate regarding Roosevelt's election date to a fellow Israeli in casual conversation but not in prepping for a U.S.-history exam, and not when the question was posed by an American acquaintance or by an Israeli historian (cf. Isaacs & Clark, 1987; Vandierendonck & Van Damme, 1988).

Part of the problem—and challenge—in understanding how informativeness and accuracy are monitored and weighted by respondents in regulating their memory reporting is that the subjective utility of both of these factors must be considered within the social-functional context in which memory retrieval and reporting takes place. Consider, for example, our proposal that in resolving the criterion conflict in cases of UK, respondents will tend to protect the informativeness criterion at the expense of the confidence criterion. This proposal was based on the following premises and logic: (1) Beyond the goal of providing accurate and useful information to others, respondents are also interested in social rewards, such as making a good impression. (2) The social rewards and penalties stemming from the informativeness of one's answers are generally experienced immediately, whereas those stemming from the correctness of one's answers are often experienced only at a later time, if at all. (3) Extrapolating from research on the discounted utility of delayed outcomes in intertemporal choice (e.g., Frederick, Loewenstein, & O'Donoghue, 2002; Read, 2004), immediate outcomes tied to the informativeness of one's answers should be weighted more heavily than delayed outcomes tied to their correctness (for a similar analysis, see Yaniv & Foster, 1995, 1997). On the basis of the present results, despite the strong tendency to violate the confidence criterion under conditions of low objective and subjective knowledge, in defiance of the original satisficing model, we cannot know whether there was in fact a greater reluctance to violate the informativeness criterion than the confidence criterion (compare Answers *C* and *E* in Figure 1B). Hence, this too becomes an interesting open issue for future memory research.

The importance of considering personal–social goals, such as impression management, that may interact with the goals of accuracy and informativeness in guiding the question answering process, has also been emphasized within the social-communication approach to memory, mentioned in the introduction of this article. Smith and Clark (1993), for example, pointed out the limitations of the traditional approach taken in memory research, in which the answering process is "viewed simply as a matter of recall," proposing instead a more social-interactive view, in which respondents "also have to deal with their self-presentation—how they look to the questioner" (p. 26). On the basis of their analyses of participants' responses to general-knowledge questions in a conversational setting (and subsequent metacognitive ratings), they concluded the following:

> Answering questions . . . is more complicated than generally supposed. Most models assume that it requires (1) memory search or activation and (2) monitoring that search . . . . In conversational settings, however, answering questions also requires a third process: (3) assessing how the response will be viewed. That process is needed to decide whether to introduce interjections, self-talk, and other commentary, what these fillers should imply, and how to express the final answer or nonanswer. Speakers keep remarkably close track of all three types of information from the moment they begin formulating their responses. How they do so must be accounted for in any adequate model of speech production. (Smith & Clark, 1993, p. 37)

Smith and Clark's (1993) remarks point to the potentially complex interplay among cognitive, metacognitive, social, motivational, and linguistic factors involved in answering questions from memory (see also Clark, 1996), presenting this as an important challenge to students of language and speech production. We would present a similar challenge to students of memory (see also Neisser, 1988), in striving to develop more complete models of memory that gradually clarify the processes by which rememberers strategically regulate their memory performance. As Neisser (1996) has eloquently argued, remembering is a form of "purposeful action," and hence, any complete theory of memory "retrieval" will need to deal with "the reason for retrieval . . . with persons, motives, and social situations" (Neisser, 1988, p. 553).

In the present research, we used a research paradigm that offers the benefits of experimental control and rigor while still tapping some of the complexity of the strategic regulation of memory performance in real-life settings. Thus, for example, we attempted to achieve a crude approximation of a real-life memory situation

by asking our participants to answer as they would when conversing with a friend, and by allowing them a great deal of control over their memory reporting relative to most memory research. Of course, asking people to mentally simulate a conversational context is not the same as actually placing them in one, and notwithstanding the control allowed over grain size and report option, the question-and-answer situation was not interactive, and the answer format was restricted to quantitative values or intervals (but see Weber & Brewer, 2008, for a generalization involving qualitative linguistic categories). These and many other deviations from real-life memory reporting may or may not limit the ecological validity of our results. Nonetheless, by securing a place for the strategic regulation of memory reporting within the memory research agenda, we hope to motivate the development of more complex and perhaps realistic paradigms that will address such issues—and beyond.

References

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

Benjamin, A. S. (2008). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. In A. Benjamin & B. Ross (Eds.), *Psychology of learning and motivation, Vol. 48: Memory use as skilled cognition* (pp. 175–223). San Diego, CA: Elsevier.

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

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. *Psychology of Learning and Motivation, 32,* 275–318.

Clark, H. H. (1996). *Using language.* New York: Cambridge University Press.

Clark, H. H., & Schober, M. F. (1992). Asking questions and influencing answers. In J. M. Tanur (Ed.), *Questions about questions* (pp. 15– 48). New York: Russell Sage Foundation.

Cruse, D. A. (1977). The pragmatics of lexical specificity. *Journal of Linguistics, 13*(2), 153–164.

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

Dodson, C. S., & Schacter, D. L. (2002). Aging and strategic retrieval processes: Reducing false memories with a distinctiveness heuristic. *Psychology and Aging, 17,* 405–415.

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

Dudukovic, N. M., Marsh, E. J., & Tversky, B. (2004). Telling a story or telling it straight: The effects of entertaining versus accurate retellings on memory. *Applied Cognitive Psychology, 18,* 125–143.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature, 40,* 351–401.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108,* 316–355.

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

Gibbs, R. W., & Bryant, G. A. (2008). Striving for optimal relevance when answering questions. *Cognition, 106,* 345–369.

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

Gilboa, A., Alain, C., Stuss, D. T., Melo, B., Miller, S., & Moscovitch, M. (2006). Mechanisms of spontaneous confabulations: A strategic retrieval account. *Brain, 129,* 1399–1414.

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

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

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

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

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

Henkel, L. A., Johnson, M. K., & De Leonardis, D. M. (1998). Aging and source monitoring: Cognitive processes and neuropsychological correlates. *Journal of Experimental Psychology: General, 127,* 251–268.

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

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

Higham, P. A., & Arnold, M. M. (2007). How many questions should I answer? Using bias profiles to estimate optimal bias and maximum score on formula-scored tests. *European Journal of Cognitive Psychology, 19,* 718–742.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin, 118,* 248–271.

Hyman, I. E. (1994). Conversational remembering: Story recall with a peer versus for an experimenter. *Applied Cognitive Psychology, 8,* 49–66.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General, 116,* 26–37.

Jacoby, L. L., Bishara, A. J., Hessels, S., & Toth, J. P. (2005). Aging, subjective experience, and cognitive control: Dramatic false remembering by older adults. *Journal of Experimental Psychology: General, 134,* 131–148.

Jost, J. T., Kruglanski, A. W., & Nelson, T. O. (1998). Social metacognition: An expansionist review. *Personality and Social Psychology Review, 2,* 137–154.

Juslin, P., Winaman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review, 107,* 384–396.

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

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

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

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence

memory: A theoretical analysis. *Journal of Memory and Language, 29,* 133–159.

Klayman, J., Soll, J. B., Gonzáles-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes, 79,* 216–247.

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

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., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General, 123,* 297–315.

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

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

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

Koriat, A., & Goldsmith, M. (n.d.) [The time course of forgetting: Focus on memory accuracy and mediating mechanisms]. Unpublished raw data.

Koriat, A., Goldsmith, M., & Halamish, V. (2008). Control processes in voluntary remembering. In H. L. Roediger III (Ed.), *Cognitive psychology of memory* (Vol. 2 of *Learning and memory: A comprehensive reference,* 4 vols. [J. Byrne, Ed.], pp. 307–324). Oxford, England: Elsevier.

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

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 34–53.

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

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

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

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology, 14,* 78–196.

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

Lindsay, D. S., Johnson, M. K., & Kwon, P. (1991). Developmental changes in memory source monitoring. *Journal of Experimental Child Psychology, 52,* 297–318.

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

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

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.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation, 26,* 125–173.

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.

Odegard, T. N., & Lampinen, J. M. (2006). Memory editing: Knowledge, criteria, and alignment. *Memory, 14,* 777–787.

Pansky, A., Goldsmith, M., Koriat, A., & Pearlman-Avnion, S. (in press). Memory accuracy in old age: Cognitive, metacognitive, and neurocognitive determinants [Special issue on aging, cognition, and neuroscience]. *European Journal of Cognitive Psychology.*

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

Perfect, T. J., & Schwartz, B. L. (Eds.). (2002). *Applied metacognition.* New York: Cambridge University Press.

Read, D. (2004). Intertemporal choice. In D. J. Koehlerand & N. Harvey (Eds.), *The Blackwell handbook of judgment and decision making* (pp. 424–443). Oxford, England: Blackwell.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology, 19,* 90–138.

Reder, L. M. (1988). Strategic control of retrieval strategies. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 22, pp. 227–259). San Diego, CA: Academic Press.

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

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

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

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

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

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 93–105.

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

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language, 32,* 25–38.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 299–314.

Sperber, D., & Wilson, D. (1995). *Relevance: Cognition and communication* (2nd ed.). Oxford, England: Blackwell.

Teigen, K. H. (1988). The language of uncertainty. *Acta Psychologica, 68,* 27–38.

van der Henst, J.-B., Carles, L., & Sperber, D. (2002). Truthfulness and relevance in telling the time. *Mind & Language, 17,* 457–466.

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

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

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

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

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

Appendix

Procedure for Criterion Estimation Based on Shape of Confidence Distribution

The aim of the described procedure was to estimate the confidence criterion used by each participant in controlling the grain size of his or her answers, on the basis of the shape of the distribution of confidence ratings associated with that participant's answers in the free-grain phase. The procedure assumes that the criterion can be identified as a particular confidence level: (a) below which there is a stable and relatively low frequency of answering, and (b) immediately above which there is a sharp increase in the frequency of answering.

The estimation procedure was performed for each participant as follows:

1. The free-grain confidence ratings for all free-grain answers, both low-knowledge (LK) and moderate-knowledge (MK), were categorized into 1 of 10 confidence categories (0–9, 10–19, 20–29, . . ., 90–100; cf. Figure 3 earlier).

2. To reduce the chance of capitalizing on spurious local fluctuations in confidence frequency, we applied a smoothing operation: The frequency of each confidence category (except for the lowest category) was replaced by the average of the frequencies of that category and the next lower adjacent category. For example, the frequency of the 10–19 category was replaced with the average of the 0–9 and 10–19 categories so that it now reflected the frequency of confidence ratings between 0 and 19.

3. The original 0–9 category was discarded, leaving nine new, smoothed (partially overlapping) categories: 0–19, 10–29, 20–39, . . ., 80–100.

4. For each of the smoothed categories (except for the two extreme categories, 0–19 and 80–100, for which the calculation was undefined), a "step" index ($step_i$) was calculated as the difference between the frequency values of this and the next higher confidence category ($confcat_{i+1} - confcat_i$) minus the absolute value of the difference between the frequency values of this and the next lower confidence category $|confcat_i - confcat_{i-1}|$.

5. The midpoint of the confidence category with the highest step index was chosen as the estimated confidence criterion for the participant. Essentially, this corresponded to the confidence level above which the slope of the (smoothed) confidence distribution was maximal, and below which it was minimal. Note that because the step index could not be calculated for the two extreme (smoothed) categories, 0–19 and 80–100, the possible criterion estimates ranged between .20 and .80 (i.e., between the 10–29 and 70–89 categories, inclusive).

The uniqueness of the chosen criterion estimate for each participant can be gauged by comparing the step index for that estimate with the next highest step estimate yielded by one of the other candidate categories. Across the three experiments using the procedure in this article, the step index corresponding to the chosen criterion estimate averaged 3.13 items, whereas the next highest step index averaged only 0.90 items. Thus, on average, the chosen criterion estimate yielded a step index that was more than 3 times higher than the next best alternative.

Received March 5, 2008
Revision received May 7, 2008
Accepted May 9, 2008 ■