BRIEF REPORT

Metacognitive effects of initial question difficulty on subsequent memory performance

Ainat Pansky · Morris Goldsmith

Abstract In two experiments, we examined whether relative retrieval fluency (the relative ease or difficulty of answering questions from memory) would be translated, via metacognitive monitoring and control processes, into an overt effect on the controlled behavior—that is, the decision whether to answer a question or abstain. Before answering a target set of multiple-choice general-knowledge questions (intermediate-difficulty questions in Exp. 1, deceptive questions in Exp. 2), the participants first answered either a set of difficult questions or a set of easy questions. For each question, they provided a forced-report answer, followed by a subjective assessment of the likelihood that their answer was correct (confidence) and by a free-report control decision—whether or not to report the answer for a potential monetary bonus (or penalty). The participants' ability to answer the target questions (forced-report proportion correct) was unaffected by the initial question difficulty. However, a predicted *metacognitive contrast effect* was observed: When the target questions were preceded by a set of difficult rather than easy questions, the participants were more confident in their answers to the target questions, and hence were more likely to report them, thus increasing the quantity of freely reported correct information. The option of free report was more beneficial after initial question difficulty than after initial question ease, in terms of both the gain in accuracy (Exp. 2) and a smaller cost in quantity (Exps. 1 and 2). These results demonstrate that changes in subjective experience can influence metacognitive monitoring and control, thereby affecting free-report memory performance independently of forced-report performance.

Keywords Relative fluency · Metacognition · Monitoring and control · Confidence · Free-report memory testing

A. Pansky (✉) · M. Goldsmith
Department of Psychology, University of Haifa,
Haifa 3498838, Israel
e-mail: pansky@research.haifa.ac.il

How do people monitor whether an answer that comes to mind is correct? An implicit metacognitive cue that is used to evaluate the accuracy of retrieved information is the ease with which the information comes to mind when attempting to retrieve it (e.g., Benjamin, Bjork, & Schwarz, 1998). This heuristic, known as *retrieval fluency*, is usually, but not always, valid (for reviews, see Benjamin & Bjork, 1996; Kelley & Rhodes, 2002). For example, Costermans, Lories, and Ansay (1992) found higher confidence in answers to general-knowledge (GK) questions that were retrieved more quickly, whether correctly or incorrectly. In a similar vein, Kelley and Lindsay (1993) found that confidence in the answers to GK questions (but not their accuracy) increased following preexposure to plausible answers. Koriat (1995) demonstrated that the feeling of knowing that people have when they search their memory for a solicited piece of information is based on the amount of partial information accessed about the target and on the ease with which it comes to mind, regardless of its accuracy. Finally, several studies have highlighted the importance of *relative* fluency as a metacognitive cue, showing that the discrepancy between experienced and expected levels of fluency is critical in influencing memory judgments and other types of evaluations (e.g., Whittlesea & Williams, 1998; see also Hansen & Wänke, 2008; Jacoby & Dallas, 1981; McCabe & Balota, 2007).

The contribution of retrieval fluency to subjective confidence gains further potential importance when one considers the role that subjective confidence plays in guiding controlled behavior (e.g., Alter & Oppenheimer, 2009; Goldsmith & Koriat, 2008; Nelson & Narens, 1990; see also Finn, 2010). Thus, for example, when people are confident that they know the answer to a question, they will generally answer it; when lacking confidence, they may prefer to respond "don't know" (e.g., Koriat & Goldsmith, 1996). Koriat and Goldsmith (1996; Goldsmith & Koriat, 2008) put forward a framework for studying the metacognitive monitoring and control processes that mediate between the retrieval of information, on the one hand, and actual, free-report performance, on the

other. In addition to the retrieval of a best-candidate answer, a *monitoring* process is used to subjectively assess the correctness of the answer (i.e., confidence), and on that basis, to decide whether it should be reported or withheld ("don't know" response). The control decision is made by setting a report criterion on the monitoring output: An answer is volunteered if the assessed probability correct passes the criterion, but is withheld otherwise. The criterion is set on the basis of implicit or explicit payoffs that reflect the perceived gain for providing correct information, relative to the cost of providing incorrect information.

One situation in which this framework may be applied is exam taking. On exams that include open-ended or essay-type questions, one must weigh the potential gain from adding more information to one's answer, against the potential cost if the information is wrong. Also, the decision whether or not to answer specific questions becomes critical on multiple-choice exams that penalize for wrong answers in order to discourage guessing (Budescu & Bar-Hillel, 1993). A host of studies have shown that one's choices regarding when and how often to guess can have a substantial impact on one's test scores. For example, on exams that use standard formula scoring (Thurstone, 1919), test takers are generally overly conservative, tending to withhold low-confidence answers that would have been better to volunteer (e.g., Higham, 2007).

Suppose that on such an exam, in deciding whether or not to venture a particular answer, a person is trying to gauge the likelihood that this answer is correct. One way in which relative retrieval fluency might bias this evaluation is by way of a *metacognitive contrast effect* (see Hansen & Wänke, 2008): The experienced ease or difficulty of answering preceding questions might be used as a standard against which the ease or difficulty of answering subsequent questions is implicitly compared. If the preceding questions were answered easily, this may decrease the experienced fluency of retrieving answers to subsequent questions, increasing their subjective difficulty and reducing one's confidence in their correctness. Conversely, if the preceding questions were difficult to answer, this may decrease the experienced difficulty of answering later questions, increasing subjective confidence. As we discussed earlier, an increase or decrease in subjective confidence is expected to have a corresponding effect on the tendency to venture the answer rather than withhold it, ultimately affecting test performance.

In the present study, the Koriat and Goldsmith (1996) framework and accompanying research methodology were used to illuminate the manner in which relative retrieval fluency might be translated, via the operation of monitoring and control processes, into an explicit effect on controlled behavior—the decision whether to answer a question or to abstain—and to examine the consequences of this effect on actual test performance.

Experiment 1

Forty GK questions were presented in a four-alternative multiple-choice format. The test included 20 items of intermediate difficulty and either ten very difficult items or ten very easy items. The intermediate-difficulty items served as the target items, whereas the other items were used to manipulate the initial level of retrieval fluency: Before answering the target questions, one group of participants first answered the difficult questions, whereas the other group first answered the easy questions.

For each multiple-choice question, a forced-reporting phase was followed by a free-report phase, on an item-by-item basis: Participants were first asked to choose one of four alternative answers and to provide a confidence judgment assessing the likelihood that their answer was correct (forced-report phase). They were then asked to decide whether they wished to volunteer the answer (free-report phase) for a potential monetary bonus, with a penalty incurred for volunteering wrong answers.

The initial-difficulty manipulation was not expected to affect the participants' ability to answer the intermediate-level target questions. However, by way of a metacognitive contrast effect, it was expected to bias the level of confidence that participants would attach to their answers, and thereby the decision whether to volunteer or withhold their answers. More specifically, we expected that answering intermediate-level questions would be experienced as relatively more fluent after answering an initial set of difficult questions than after answering an initial set of easy questions, resulting in higher confidence in the answers, and consequently, in a higher volunteering rate of answers in the former than in the latter condition. Given that each volunteering decision was reported immediately after the confidence judgment, it was possible that this decision might be affected by the explicit reporting of confidence. In order to make sure that the explicit solicitation of confidence judgments did not bias any effects of initial question difficulty, we collected confidence judgments for only half of the participants in each group, and examined whether this report condition (whether or not confidence judgments were solicited) interacted with initial question difficulty in affecting any of the dependent measures.

Method

Participants A group of sixty-four undergraduates from the University of Haifa participated in the experiment. They were randomly and equally assigned to each of the experimental groups.

Materials A subset of 40 GK questions were selected from 60 four-alternative multiple-choice recognition questions (in Hebrew) on the basis of preliminary testing with 20 participants. Following Koriat (1995), the assessment of question difficulty

was based on the proportion of participants who freely provided an answer to each question. The ten questions with the lowest tendency to be freely reported ($M = .07$) were selected for the initially difficult condition, whereas the ten questions with the highest tendency to be freely reported ($M = .69$) were selected for the initially easy condition. Finally, the 20 questions in the center of the volunteering rate distribution ($M = .36$) were selected as the target questions of intermediate difficulty. From these questions, two versions of the GK test were compiled. Both versions contained the 20 intermediate-difficulty target questions. However, in the initially difficult version, these questions were preceded by the ten difficult questions, whereas in the initially easy version, they were preceded by the ten easy questions. Finally, the order within each set of questions (initially difficult or initially easy, plus intermediate difficulty) was counterbalanced across participants, creating four variations of each test version.

Procedure Half of the participants in each group (initially difficult or initially easy) performed the following procedure for each test question. They were first asked to select one of the four alternatives that they thought was the correct answer. They were required to answer all of the questions, even if they had to guess. Next, they were required to assess the likelihood that their answer was correct, using a 25 %–100 % scale. No monetary incentive was offered for performance on these two tasks, which constituted the forced-report phase. Finally, in the free-report phase, the participants were asked to decide whether to volunteer or withhold their response. Volunteering accurate responses was induced by a moderate-incentive payoff schedule: For each volunteered answer, the participant gained one point if it was correct, but lost one point if it was incorrect. The participants were told that they would not be penalized (but neither would they receive any bonus) for withheld responses, and that their score would be translated to a monetary bonus. This three-step procedure (i.e., forced-report response, confidence judgment, volunteer decision) was repeated for each of the 30 GK questions on an item-by-item basis. The second half of participants in each group performed the same procedure for each test question, except that these participants were not requested to provide a confidence judgment.

Results and discussion

First, as a manipulation check, the difficulty of the two sets of initial questions was compared. As expected, the proportion of correct responses was much lower for the questions assigned to the initially difficult condition (.30) than for those assigned to the initially easy condition (.76), $t(62) = 8.55$, $p < .001$, $d = 2.17$. The initially difficult questions were also subjectively experienced as more difficult than the initially easy questions, with substantially lower confidence judgments associated with the answers to the former (44 %) than with the answers to the latter (71 %), $t(30) = 5.61$, $p < .001$, $d = 2.05$. Finally, answers to the initially difficult questions were less likely to be freely volunteered (.22) than answers to the initially easy questions (.62), $t(62) = 7.69$, $p < .001$, $d = 1.92$.

In order to test our hypotheses regarding the effect of initial question difficulty on subsequent responding to the intermediate-difficulty target questions, we subjected each of the dependent measures (i.e., forced-report proportion correct, mean confidence, volunteering rate, free-report memory quantity, free-report memory accuracy, and total points achieved) to a between-subjects analysis of variance (ANOVA). In these analyses, initial question difficulty (initially easy or initially difficult) and report condition (whether or not confidence judgments were solicited) served as independent variables, and the presentation order of the initial questions and presentation order of the target questions served as independent control variables. We found that neither of these control variables interacted with the difficulty of the manipulation questions. Therefore, the data were pooled across the various question orders. We also found that whether or not an explicit confidence judgment was solicited did not interact with initial difficulty. The means and standard deviations of the dependent measures for each of the experimental groups are presented in Table 1.

First, we expected that initial question difficulty would not affect the ability to answer the intermediate-difficulty target questions, and, indeed, the forced-report proportions correct were comparable for the participants who had initially

Table 1 Performance measures and metacognitive indices for intermediate-difficulty questions, following an initial set of easy versus difficult questions (Exp. 1)

Initial Difficulty	Forced-Report Proportion Correct	Confidence	Volunteering Rate	Free-Report Memory Quantity	Free-Report Memory Accuracy	Total Points Achieved	Report Criterion
Easy	.52 (.14)	.52 (.13)	.30 (.17)	.23 (.14)	.73 (.27)	3.00 (3.42)	.69 (.16)
Difficult	.55 (.16)	.64 (.16)	.50 (.19)	.35 (.17)	.71 (.22)	4.13 (4.38)	.62 (.27)

Standard deviations appear in parentheses

answered difficult questions (.55) as for those who had initially answered easy questions (.52), $F < 1$, $\eta^2_p = .01$.

Second, as expected, initial question difficulty did affect the subjective confidence associated with the responses to the target questions (expressed as the assessed probability correct, ranging between .25 and 1): The participants who had initially answered difficult questions were more confident in their answers to the subsequent target questions (.64) than were those who had initially answered easy questions (.52), $F(1, 30) = 5.05$, $p = .032$, $\eta^2_p = .15$. As a result, whereas participants in the initially easy group were perfectly calibrated in assessing the correctness of the answers to the target questions, exhibiting no difference between the estimated probability of correctness and the actual proportion correct, $t < 1$, $d = 0.10$, participants in the initially difficult group were overconfident (by .09), $t(15) = 3.23$, $p = .006$, $d = 0.81$. The difference in overconfidence between the two groups was marginally significant, $t(30) = 2.00$, $p = .055$, $d = 0.73$.

Third, the effect of initial question difficulty on confidence also translated into an overt effect on controlled behavior, with a higher tendency to volunteer an answer after initially answering difficult questions (.50) than after initially answering easy questions (.30), $F(1, 60) = 18.30$, $p < .001$, $\eta^2_p = .23$. Importantly, as we noted earlier, the same pattern was obtained whether or not confidence judgments were explicitly collected, with a nonsignificant interaction between initial difficulty and report condition, $F < 1$, $\eta^2_p = .01$.

Next, we examined the extent to which the effect of initial question difficulty on confidence, and consequently on volunteering rate, also affected free-report memory performance on the target questions. Indeed, initial question difficulty affected the *quantity* of correct information provided under free report, such that the proportion of correct volunteered answers (out of the total number of test questions) was higher for the initially difficult group (.35) than for the initially easy group (.23), $F(1, 60) = 9.84$, $p = .003$, $\eta^2_p = .14$. Free-report *accuracy*, however, was not affected by initial question difficulty, with comparable proportions of correct answers out of the total number of volunteered answers for the initially difficult group (.71) and the initially easy group (.73), $F < 1$, $\eta^2_p = .003$. Given the higher memory quantity with no difference in accuracy, one would expect to see an advantage for the initially difficult group in terms of the number of points gained under the operative scoring rule. Although the observed difference was in that direction (4.13 and 3.00 for the initially difficult and initially easy groups, respectively), it did not reach statistical significance, $F(1, 60) = 1.35$, $p = .25$, $\eta^2_p = .02$.

As is shown in Fig. 1a, the option of free report allowed both groups to achieve comparable and significant gains in accuracy (of .19, on average) relative to the forced-report condition, $F(1, 62) = 48.60$, $p < .001$, $\eta^2_p = .44$, with a nonsignificant interaction between report option (forced or

free) and initial difficulty, $F(1, 62) = 1.03$, $p = .31$, $\eta^2_p = .02$. However, the gain in accuracy with the exercise of free report came at a cost in quantity, $F(1, 62) = 314.85$, $p < .001$, $\eta^2_p = .84$, evidencing a quantity–accuracy trade-off. This cost was smaller for the initially difficult group (.20) than for the initially easy group (.29), with a significant interaction between report option and group, $F(1, 62) = 12.11$, $p = .001$, $\eta^2_p = .16$. In other words, the initially difficult group exhibited a milder quantity–accuracy trade-off than did the initially easy group.

Finally, we examined whether initial question difficulty had an effect on one's control policy, in order to consider the possibility that the higher volunteering rate and free-report quantity that we found after answering difficult questions was due to a more liberal control policy. To estimate each participant's criterion level, we used a computational procedure developed by Koriat and Goldsmith (1996). Considering each probability rating actually used by the participant as a candidate P_{rc} (criterion), hits were defined as volunteered answers for which P_a (assessed probability) $\geq P_{rc}$, and correct rejections as withheld answers for which $P_a < P_{rc}$. False alarms and misses were also defined accordingly. The chosen P_{rc} estimate for each participant was the value that maximized the percentage of hits and correct rejections combined (*fit rate* averaging 92 % across participants). The mean estimated P_{rc} value for the initially difficult group (.62) was somewhat lower than that for the initially easy group (.69), but the difference was not significant, $t < 1$, $d = 0.33$. In addition, the effects of initial difficulty on volunteering rate and free-report quantity remained significant when report criterion was added as a covariate, $F(1, 29) = 8.49$, $p = .007$, $\eta^2_p = .23$, and $F(1, 29) = 5.13$, $p = .031$, $\eta^2_p = .15$, respectively, suggesting that the difference between the groups in volunteering rates and free-report quantity performance stemmed primarily from an effect of initial difficulty on control that was mediated by monitoring (i.e., subjective confidence), rather than from a direct effect on the placement of the report criterion (but see the General Discussion).

To summarize the findings, when the target questions were preceded by a set of difficult rather than easy questions, participants were more confident in their answers to the target questions, and hence were more likely to report them, thus increasing the quantity of freely reported correct information.

Experiment 2

Previous research has shown that with increasing monitoring difficulty, the option of free report becomes less beneficial for memory accuracy, and the cost of withholding answers becomes relatively high in terms of memory quantity (e.g., Koriat & Goldsmith, 1996). In Experiment 2, the effects of initial difficulty on free-report memory performance were examined under more difficult monitoring conditions, in an attempt to detect differences in free-report accuracy, in
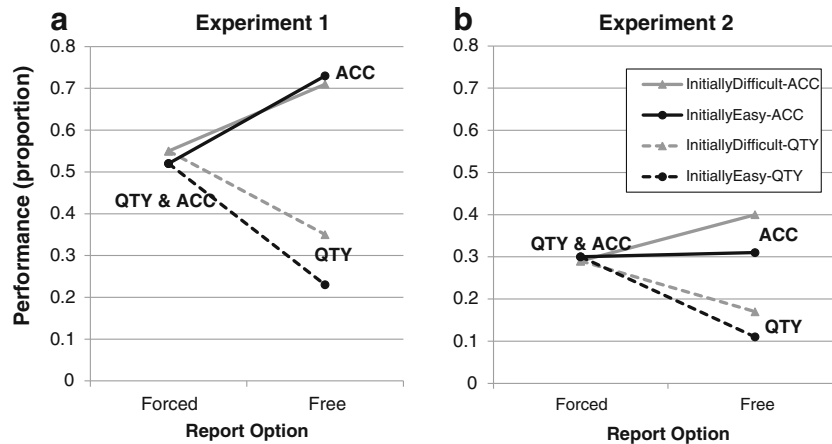
Fig. 1 Quantity (QTY: i.e., the proportion of questions answered correctly) and accuracy (ACC: i.e., the proportion of reported answers that were correct) under free versus forced memory reporting, following an initial set of either easy or difficult questions, in Experiment 1 (**a**) and

Experiment 2 (**b**). Note that quantity and accuracy measures can be distinguished operationally only under conditions of free report, whereas these measures are operationally equivalent under forced-report conditions, in which all questions must be answered

addition to replicating the effects of initial difficulty on free-report memory quantity (obtained in Exp. 1). In Experiment 2, the intermediate-difficulty target questions were replaced with "deceptive" questions—items for which the participants' ability to monitor the correctness of their answers was expected to be poor and overconfidence was expected to be high (see Fischhoff, Slovic, & Lichtenstein, 1977; Koriat, 1995).

Method

Participants A group of sixty-four undergraduates from the University of Haifa, who had not participated in Experiment 1, took part in this experiment. They were randomly and equally assigned to each of the experimental groups.

Materials The materials were identical to those used in Experiment 1, except that the 20 intermediate-difficulty target questions were replaced by 20 "deceptive" questions. On the basis of preliminary testing with ten participants, subjective confidence averaged .60 for these items, whereas the forced-report proportion correct averaged only .30, yielding pronounced overconfidence. Given that neither the presentation order of the initial questions nor the presentation order of the target questions had any effect in Experiment 1, only one presentation order was used for each group.

Procedure The procedure was identical to that of Experiment 1.

Results and discussion

The analyses conducted on the data of Experiment 1 were repeated for those of Experiment 2. As expected, and as before, the initially difficult questions were inferior to the initially easy questions in terms of proportion correct (.33 vs. .76, respectively), $t(62) = 9.46$, $p < .001$, $d = 2.40$, confidence (39.34 % vs. 71.66 %, respectively), $t(30) = 8.89$, $p < .001$, $d = 3.25$, and volunteering rate (.30 vs. .60, respectively), $t(62) = 5.25$, $p < .001$, $d = 1.33$.

As in Experiment 1, whether or not an explicit confidence judgment was solicited did not interact with initial difficulty for any dependent measure. The mean and standard deviation of each dependent measure, calculated for the target deceptive questions for each of the experimental groups, are presented in Table 2. Again, equivalent levels of forced-report proportions correct were observed for the participants who had initially answered difficult questions (.29) and those who had initially answered easy questions (.30), $F < 1$, $\eta^2_p < .001$. In line with our predictions, the assessed probability correct of the answers was higher for the initially difficult group (.57) than for the initially easy group (.50), but the difference only approached statistical significance, $F(1, 30) = 2.56$, $p = .12$, $\eta^2_p = .08$. Also, an observed numerical difference in overconfidence in

Table 2 Performance measures and metacognitive indices for deceptive questions, following an initial set of easy versus difficult questions (Exp. 2)

Initial Difficulty	Forced-Report Proportion Correct	Confidence	Volunteering Rate	Free-Report Memory Quantity	Free-Report Memory Accuracy	Total Points Achieved	Report Criterion
Easy	.30 (.17)	.50 (.13)	.33 (.20)	.11 (.13)	.31 (.27)	–2.06 (3.61)	.62 (.18)
Difficult	.29 (.10)	.57 (.13)	.49 (.27)	.17 (.10)	.40 (.22)	–2.91 (4.23)	.60 (.21)

Standard deviations appear in parentheses

the same direction (overconfidence = .28 and .20, respectively) was not significant, $t(30) = 1.19$, $p = .242$, $d = 0.43$.

As in Experiment 1, the volunteering rate for the target questions was substantially higher after initially answering difficult questions (.49) than after initially answering easy questions (.33), $F(1, 60) = 7.04$, $p < .001$, $\eta^2_p = .11$. As before, we observed no difference in the estimated P_{rc}s between the initially difficult group (.60) and the initially easy group (.62), $t(30) = 0.9$, $p = .777$, $d = 0.33$ (fit rate: 94 %). Once again, this implies that the effect of initial difficulty was mediated by confidence rather than by report criterion. Yet, the earlier-reported effect on mean confidence only approached statistical reliability. To resolve this puzzle, we examined whether the effect of initial question difficulty on confidence might be expressed as a distribution shift, affecting the proportion of answers that were assigned medium-to-high levels of confidence—presumably those answers that were actually considered by the participant as plausible candidates for volunteering. Operationally, we defined medium-to-high confidence answers as those with an assessed probability correct higher than one standard deviation below the mean P_{rc} (assessed probability correct > .42). In line with the predicted effect of initial question difficulty on subjective confidence, and with its observed effect on volunteering rate, the proportion of such medium-to-high confidence answers was higher in the initially difficult group (.64) than in the initially easy group (.48), $t(30) = 1.78$, $p = .09$, $d = 0.65$ (significant by a one-tailed test).

The effect of initial question difficulty on confidence, and consequently on the volunteering rate, also affected free-report quantity performance, such that the proportion of correct volunteered answers (out of the total number of test questions) was higher for the initially difficult group (.17) than for the initially easy group (.11), $F(1, 60) = 4.17$, $p = .046$, $\eta^2_p = .0.07$. Free-report memory accuracy was numerically higher for the initially difficult group (.40) than for the initially easy group (.31), but this difference only approached statistical significance, $F(1, 60) = 2.32$, $p = .133$, $\eta^2_p = .04$; we observed no difference in points gained, $F < 1$, $\eta^2_p = .01$. Nevertheless, as is shown in Fig. 1b, when they were allowed the option of free report, the initially difficult group achieved a reliable gain in accuracy (.11), $t(31) = 2.89$, $p = .007$, $d = 1.04$, whereas the initially easy group did not (.01), $t(31) = 0.34$, $p = .733$, $d = 0.12$ [$F(1, 62) = 3.75$, $p = .06$, $\eta^2_p = .06$, for the interaction between report option (forced or free) and group]. The initially difficult group also exhibited a smaller decrease in quantity when utilizing the option of free report (.12) than did the initially easy group (.18), $F(1, 62) = 6.21$, $p = .02$, $\eta^2_p = .09$. Thus, under the difficult monitoring conditions of Experiment 2, the initially difficult group exhibited a quantity–accuracy trade-off similar to the one observed in Experiment 1, whereas the initially easy group not only paid a price in quantity, but also failed to improve at all in accuracy when exercising the free-report option.

General discussion

In the present study, we manipulated the difficulty of an initial set of questions and examined its effect on subsequent responding. The participants' ability to answer the target questions (i.e., their forced-report proportions correct) was unaffected by initial question difficulty. However, the predicted metacognitive contrast effect was observed: When the target questions were preceded by a set of difficult rather than easy questions, the participants were more confident in their answers to the target questions, and hence were more likely to report them, thereby increasing the quantity of correct information that was volunteered (without a reduction in accuracy). Importantly, the higher volunteering rate and higher amount of correct freely reported information that we found after answering difficult questions was obtained whether or not confidence judgments were explicitly solicited, ruling out a demand characteristic interpretation that would attribute these findings to an artificially high correlation between the reported subjective measure and actual behavior.

A plausible interpretation for the higher volunteering rate and free-report quantity among the participants in the initially-difficult group is that, after volunteering fewer answers in the first stage than the participants in the initially-easy group, they compensated by volunteering more answers in the following stage. Alternatively, due to the greater difficulty of the test in the initially difficult group, these participants may have set a more liberal report criterion to begin with, which remained stable even after the questions became easier. To discount such criterion-based hypotheses, we estimated the report criterion used on the intermediate-difficulty items on the basis of the confidence ratings and volunteering decisions, and we found that the criteria did not differ between the two groups. One should note, however, that the confidence ratings used to estimate the report criterion might themselves have been biased by criterion placement. Distinguishing between effects on "true" subjective confidence versus effects on the mapping between subjective confidence and confidence ratings is notoriously difficult (for a related discussion, see Goldsmith, 2011; Higham, 2011). Thus, the present findings cannot be treated as definitive evidence for a metacognitive contrast effect.

Nevertheless, we have shown that initial item difficulty yields a contrast effect on confidence judgments, accompanied by a corresponding effect on the reporting rate—joint effects that were predicted to follow from a contrast effect on subjective confidence. In contrast, we see no reason to expect that initial question difficulty would affect the mapping between subjective confidence and the numbers used to indicate that confidence, particularly given that the direction of the effect would have been opposite to what is expected on the basis of "anchoring and adjustment" (Tversky & Kahneman,

1974). Thus, we continue to interpret the observed effects on reporting behavior as being mediated by subjective experience, though further examination of this issue is called for.

Previous findings obtained by Weinstein and Roediger (2010, 2012) also demonstrated the effect of initial difficulty on subjective experience. The participants in these studies believed that they had answered more questions correctly when the questions were sorted from the easiest to the hardest than when the questions were randomized or sorted from the hardest to the easiest, presumably because the difficulty of the initial questions anchored participants' evaluations of performance throughout the remainder of the test. In a similar vein, answering an initial set of easy questions in the present study may have anchored an expectation for relative ease of retrieval. The much more difficult answers that came later were then experienced as unexpectedly difficult, thereby yielding lower confidence, volunteering rate, and free-report quantity. In contrast to the present findings, Weinstein and Roediger (2010, 2012) did not find any effect of initial question difficulty on the item-by-item ratings, perhaps because the change in difficulty in their studies was gradual, in contrast to the abrupt change in our study. This differential pattern is compatible with earlier findings suggesting that the influence of a preceding judgment depends on the perceived similarity between the current target and its predecessor: Assimilation effects are observed when a target is perceived as being generally similar to its predecessor, whereas contrast effects are observed when it is perceived as being dissimilar (e.g., Damisch, Mussweiler, & Plessner, 2006).

Bodner and Richardson-Champion (2007) found higher recognition rates and more "remember" (rather than "know") judgments for medium-difficulty details from a crime film following a block of difficult-to-retrieve details than following a block of easy-to-retrieve details. It would be interesting in future research to apply the present paradigm to an eyewitness situation and demonstrate the effect of initial difficulty on free-report eyewitness reporting via its effect on subjective experience.

In conclusion, our results demonstrate that changes in subjective experience can influence metacognitive monitoring and control, thereby affecting free-report memory performance independently of forced-report performance. More concretely, they show that initial item difficulty can affect the joint levels of free-report quantity and accuracy that are subsequently achieved, via its influence on the underlying metacognitive processes.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13,* 219–235.

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

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

Bodner, G. E., & Richardson-Champion, D. D. L. (2007). Remembering is in the details: Effects of test-list context on memory for an event. *Memory, 15,* 718–729.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement, 30,* 277–291.

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

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgements. *Journal of Experimental Psychology: Applied, 12,* 166–178.

Finn, B. (2010). Ending on a high note: Adding a better end to effortful study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1548–1553.

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

Goldsmith, M. (2011). Quantity-accuracy profiles or type-2 signal detection measures? Similar methods toward a common goal. In P. A. Higham & J. P. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea* (pp. 128–136). Basingstoke, UK: Palgrave-Macmillan.

Goldsmith, M., & Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation* (Vol. 48, pp. 1–60). London, UK: Academic Press.

Hansen, J., & Wänke, M. (2008). It's the difference that counts: Expectancy/experience discrepancy moderates the use of ease of retrieval in attitude judgments. *Social Cognition, 26,* 447–468.

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. (2011). Accuracy discrimination and type-2 signal detection theory: Clarifications, extensions, and an analysis of bias. In P. A. Higham & J. P. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea* (pp. 109–127). Basingstoke, UK: Palgrave-Macmillan.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306–340.

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

Kelley, C. M., & Rhodes, M. G. (2002). Making sense and nonsense of experience: Attributions in memory and judgment. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 41, pp. 293–320). San Diego, CA: Academic Press.

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

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

McCabe, D. P., & Balota, D. A. (2007). Context effects on remembering and knowing: The expectancy heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 536–549.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125–173). San Diego, CA: Academic Press.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Weinstein, Y., & Roediger, H. L., III. (2010). Retrospective bias in test performance: Providing easy items at the beginning of a test makes students believe they did better on it. *Memory & Cognition, 38,* 366–376.

Weinstein, Y., & Roediger, H. L., III. (2012). The effect of question order on evaluations of test performance: How does the bias evolve? *Memory & Cognition, 40,* 727–735.

Whittlesea, B. W. A., & Williams, L. D. (1998). Why do strangers feel familiar, but friends don't? A discrepancy-attribution account of feelings of familiarity. *Acta Psychologica, 98,* 141–165.