Journal of Experimental Psychology: Applied, 2011, Vol. 17, No. 2, 148 –158

Regulation of Memory Accuracy With Multiple Answers: The Plurality Option

Karlos Luna, University of Minho

Philip A. Higham, University of Southampton

Beatriz Martín-Luengo, University of Granada

Karlos Luna, School of Psychology, University of Minho, Braga, Portugal; Philip A. Higham, School of Psychology, University of Southampton, Southampton, England; Beatriz Martíin-Luengo, School of Psychology, University of Granada, Granada, Spain.

Beatriz Martın-Luengo is now at the Faculty of Psychology, University of the Basque Country, Basque Country.

Portions of this research were presented at the 51st annual meeting of the Psychonomic Society, St. Louis, Missouri, November 18 –21, 2010.

Correspondence concerning this article should be addressed to Karlos Luna, School of Psychology, University of Minho, Campus de Gualtar, 4710-057, Braga, Portugal. E-mail: karlos.luna@psi.uminho.pt

We report two experiments that investigated the regulation of memory accuracy with a new regulatory mechanism: the plurality option. This mechanism is closely related to the grain-size option but involves control over the number of alternatives contained in an answer rather than the quantitative boundaries of a single answer. Participants were presented with a slideshow depicting a robbery (Experiment 1) or a murder (Experiment 2), and their memory was tested with five-alternative multiple-choice questions. For each question, participants were asked to generate two answers: a single answer consisting of one alternative and a plural answer consisting of the single answer and two other alternatives. Each answer was rated for confidence (Experiment 1) or for the likelihood of being correct (Experiment 2), and one of the answers was selected for reporting. Results showed that participants used the plurality option to regulate accuracy, selecting single answers when their accuracy and confidence were high, but opting for plural answers when they were low. Although accuracy was higher for selected plural than for selected single answers, the opposite pattern was evident for confidence or likelihood ratings. This dissociation between confidence and accuracy for selected answers was the result of marked overconfidence in single answers coupled with underconfidence in plural answers. We hypothesize that these results can be attributed to overly dichotomous metacognitive beliefs about personal knowledge states that cause subjective confidence to be extreme.

Keywords: regulation of memory accuracy, plurality option, dual-criterion model, metamemory, eyewitness memory, confidence–accuracy dissociation

Recent work in cognitive psychology has examined how people regulate the accuracy of their memory reports. This work has focused on two regulatory mechanisms. The first mechanism is a report option that can be used to filter out unwanted, low-quality information (e.g., Higham, 2007; Koriat & Goldsmith, 1996). For example, an eyewitness in a situation that calls for high

accuracy (e.g., giving testimony during a trial) may choose to report only details from memory that are associated with high confidence and are likely to be accurate, while saying "I don't know" to other questions. The second mechanism is a grain-size option that can be used to increase accuracy by reducing the amount of detail in memory reports (e.g., Ackerman & Goldsmith, 2008; Goldsmith Koriat, & Weinberg-Eliezer, 2002; Weber & Brewer, 2008; Yaniv & Foster, 1995, 1997). For example, when asked about the timing of a robbery, an unsure eyewitness may report "sometime between 9 p.m. and 11 p.m." rather than "10 p.m.," allowing for a margin of error and increasing the likelihood that the answer will be correct. The accuracy benefits of using either a report option or a grain-size option both come at the cost of informativeness: Coarse answers and withheld answers convey little information and no information, respectively. Therefore, excessive use of either option runs the risk of violating typical social discourse norms. To communicate effectively, people must weigh the competing goals of informativeness against accuracy, resulting in an accuracy– informativeness trade-off (see Goldsmith & Koriat, 2008, for a review).

In the current research, we investigated an accuracy regulation mechanism closely related to the grain-size option that we call the *plurality option*. This option involves control over *the number of alternatives* being offered in an answer rather than the grain size of a single answer as in previous grain-size research. For example, when asked what a suspect was wearing on his head during a robbery, an uncertain eyewitness might offer three distinct possibilities such as "a beret, a baseball cap, or a bandanna" rather than just one alternative. We consider the plurality option to be a specific instantiation of the grain-size mechanism because, regard- less of whether several different alternatives are offered or whether a coarse-grained version of a single answer is offered, both augment accuracy by reducing specificity (i.e., both options result in an accuracy–informativeness trade-off). However, as will become clear, the plurality option can be used in many circumstances in which single-answer grain-size control is not feasible. Consequently, we believe that distinguishing between the two options is important. We return to the similarities and differences between single-answer grain-size control versus plurality control in the General Discussion.

Although this research marks the first systematic investigation of the plurality option, a number of reports have been published on the control of single-answer grain size to regulate memory accuracy. Because of the close relationship between the two options, we first review research on single-answer grain-size control.

Research on Grain-Size Control

The roots of the research on grain size go back to a pair of studies by Yaniv and Foster (1995, 1997), in which the accuracy– informativeness trade-off was investigated for the first time. Participants were presented with a question, always pertaining to a numerical estimate, and had to select which of two provided intervals was the best answer (Yaniv & Foster, 1995) or provide their own interval estimates in different conditions (Yaniv & Foster, 1997). The results suggested that both accuracy and informativeness were considered in the decision process.

However, the main work on the grain-size topic was conducted after Goldsmith et al. (2002; see also Ackerman & Goldsmith, 2008; Goldsmith & Koriat, 2008; Goldsmith, Koriat, & Pansky, 2005) developed a two-phase procedure to study this trade-off using general-knowledge questions requiring numeric answers. In Phase 1, participants were asked questions such as "When did Boris Becker last win the Wimbledon men's tennis final?" Participants were asked

to provide two different answers, one fine grained (a 3-year interval) and another coarse grained (a 10-year interval). In Phase 2, participants selected which of these two answers, either fine grained or coarse grained, they would prefer to provide if they were an expert witness on a government committee. Goldsmith et al. compared overall accuracy for coarse- versus fine-grained answers in Phase 1 with accuracy for the selected answers in Phase 2. This procedure and analogous materials have been used in almost all research on this topic.

If participants can trade off informativeness for accuracy by controlling grain size, we would expect that accuracy would be higher for fine-grained answers that were selected in Phase 2 than for fine-grained answers that were rejected via selection of coarse- grained answers. A data pattern such as this one would suggest that participants attempted to reach a happy medium between accuracy and informativeness, volunteering fine-grained answers when their likelihood of accuracy was high and coarse-grained responses when their likelihood of accuracy was lower. Different variations of this basic procedure were tested by Goldsmith and others (e.g., Goldsmith et al., 2002, 2005; Weber & Brewer, 2008), who found that participants can typically regulate accuracy reasonably well by controlling the grain size of their reported answer, although the regulation was far from perfect.

Ackerman and Goldsmith (2008) developed a dual-criterion model to formalize the findings in the literature on single-answer grain-size control. In the dual-criterion model, an answer should meet two criteria to be volunteered. The confidence criterion is the minimum confidence that an answer should have to be considered high enough to be volunteered. The informativeness criterion is the degree of precision or coarseness the answer should have to be informative enough to be volunteered.

Grain-Size Control in Applied Settings

Apart from the theoretical work on the grain-size regulatory mechanism, some studies have also attempted to generalize those results to an applied setting, specifically to eyewitness testimony. Such an application could help in developing new ways to question eyewitnesses so that answers are offered with the highest possible accuracy and informativeness. In this vein, Goldsmith et al. (2005) had participants read a police interview of witnesses to an argument that took place in a pub, leading to a later assault. Participants then answered several questions about the interview, but the questions always pertained to numeric values (e.g., age, height, distance, time). The results again showed that participants regulated accuracy by controlling the grain size of their answers. Similarly, Weber and Brewer (2008) investigated the control of grain size in an eyewitness setting and extended Goldsmith et al.'s results to verbal information, but only of a specific kind: colors. In this last case, the fine-grained answers were defined as the specific color and the coarse-grained answers were defined as the overall tone (light or dark).

Although knowing how participants can regulate the accuracy of answers pertaining to colors can be as interesting in a forensic setting as the study of the regulation of accuracy pertaining to numbers, it is clear that not all the useful information a witness can provide falls into those two categories. As noted earlier, the plurality option offers an alternative to the typical instantiation of grain-size control that is not subject to these limitations because it can be used with many different kinds of questions and answers.

Overview of the Experiments

To investigate how people might regulate accuracy using plurality, we used a multiple-choice testing procedure. One aspect of this procedure, explained later, is a partial adaptation of the Subset Selection Testing used in psychometric testing (e.g., see Ben- Simon, Budescu, & Nevo, 1997; Frary, 1989).[1] In Subset Selection Testing, participants in a multiple-choice test are allowed to select as many options as they feel are necessary to include the correct answer. However, the research on Subset Selection Testing is focused on the psychometric properties of different kinds of multiple-choice tests and not on the metacognitive processes and accuracy advantage that can be gained by giving examinees control over the number of alternatives to select. To our knowledge, our research is the first to adapt Subset Selection Testing to investigating accuracy regulation.

We report two experiments in which participants first watched a video of a simulated crime scene and then completed a multiple- choice memory test about the scene. Each question on the test included five alternatives, of which only one was correct. Analogous to previous grain-size research, participants were required to generate two answers to each question using the alternatives that were provided. The first (singular) answer was analogous to a fine-grained answer as defined in previous research; it consists of one alternative (from among the five that are offered) that the participant considered most likely to be correct, guessing if necessary. The second (plural) answer was analogous to a coarse- grained answer; to generate it, participants were required to choose the second and third most plausible alternatives to add to their singular answer, also guessing if necessary (three alternatives in total). Participants were then required to rate both the singular and the plural answer on their confidence (0%–100%) that each included the correct alternative. Note that, logically, the accuracy of the plural answer must be as high or higher than that of the singular answer, and confidence should reflect that. Finally, participants chose which of those two answers, single or plural, they would like to volunteer.

In Experiment 1, the plurality option was examined directly for the first time. In Experiment 2, the results of Experiment 1 were replicated with different materials and slightly different instructions to eliminate some potentially uninteresting accounts of the data.

Experiment 1

Method

Participants. Twenty-four participants (13 men and 11 women, mean age = 24.29 years, *SD* = 3.70, range = 19 –33) from the University of Southampton (Southampton, England) took part in this experiment, in exchange for either course credits or £5 ($8.12).

Materials. We used a slideshow from Luna and Migueles (2005) that included 21 slides about a small robbery on a university campus. This slideshow depicted two friends coming out of a school and sitting on the grass. A boy approaches them to ask for a light and steals a mobile phone. After a brief scuffle between the robber and one of the friends, the robber runs away. No audio track was included in the slideshow, but expressions and gestures were explicit enough for participants to understand what was happening (see the example in the next paragraph).

For testing, we created 20 questions, each with five alternative answers. Questions followed the chronological order of the events, and the position of the alternatives (1–5) was determined by chance. All the alternatives were plausible and congruent with the event. For instance, one question was, “When the robber approached the girl, what did he do?” with the correct alternative, “Ask for a light,” and four incorrect alternatives, “Say hello,” “Ask for a pen,” “Ask for the time,” and “Ask for a cigarette.” The corresponding slide depicted the robber in front of the girl, with one arm extended and a cigarette clearly visible in his hand, in the universal gesture of asking for a light. The questions queried a variety of details such as the shape of a window, a vehicle that passed behind the two friends, and the stolen object.

Procedure. Participants were tested individually. They entered the laboratory and sat in front of a computer. The experimenter informed participants that they were to watch a slideshow about a crime that would be presented on the computer monitor, but the upcoming memory test was not mentioned. The first slide prompted the participants to press a key when they were ready, and then the slideshow began. Each slide was presented for 2 s, separated by a black screen that was shown for 1 s. When the 21 slides depicting the robbery had finished, a final slide prompted the participants to call the experimenter.

Participants then engaged in an unrelated filler task (Sudoku) for 7 min and then answered the 20 questions about the slides. First, a screen with instructions appeared on the computer monitor. The instructions indicated that participants’ memory of the slideshow would be tested and that individual questions would be presented along with five alternatives, only one of which was correct. For each question, participants were to create two answers: a single answer and a plural answer. To create the single answer, participants typed the number corresponding to one alternative that they considered correct in a text box. Just below that text box were 11 radio buttons labeled with values ranging from 0 to 100 and increasing by 10 to represent a percentage scale. Participants were asked to choose one to indicate their confidence that the answer was correct. To create the plural answer, participants typed in three different text boxes the numbers corresponding to three alternatives that they considered would include the correct alternative and again rated confidence by selecting a radio button on a second 11-point confidence scale. If the plural answer did not include the single answer, a pop-up window reminded them that it must. The order of presentation for single and plural answers was counter- balanced so that half of participants were required to create the single answer first (for all questions), and half were required to create the plural answer first (for all questions). Analysis of the presentation order did not show any significant differences, so we collapsed the data across these counterbalancing conditions. Finally, participants used a radio button to select which of these two answers (either the one-alternative single answer or the three- alternative plural answer) they would choose to report if they were a real eyewitness in a court. Questions and corresponding alternatives were presented two at a time on the computer screen. After answering all 20 questions, participants were debriefed and dis- missed. The experimental session lasted approximately 20 minutes.

Results

Participants selected the single and plural answers 43% and 57% of the time, respectively. Descriptive statistics for accuracy and confidence for all the single answers, all the plural answers, and all the selected answers can be seen in Table 1.

Mean accuracy. Mean accuracy is shown in the top panel of Figure 1. A 2 (answer type: single or plural) × 2 (decision: selected or rejected) repeated-measures analysis of variance (ANOVA)

was conducted on these data. Both the main effects of answer type, $F(1, 23) = 150.35$, $MSE = 0.02$, $p < .001$, $h^2 = .87$, and decision, $F(1, 23) = 7.85$, $MSE = 0.01$, $p = .01$, $h^2 = .25$, were significant. Accuracy was higher for plural answers ($M = .78$, $SEM = .04$) than for single answers ($M = .45$, $SEM = .05$) and also higher for selected answers ($M = .65$, $SEM = .04$) than for rejected answers ($M = .58$, $SEM = .07$). However, both of these main effects were qualified by a significant interaction, $F(1, 23) = 19.45$, $MSE = 0.05$, $p < .001$, $h^2 = .46$.

To examine this interaction in more detail, we conducted some follow-up tests. First, we tested for evidence of strategic regulation of accuracy with the plurality option. We would find such evidence

Table 1
Mean Accuracy and Confidence in Experiments 1 and 2

	Answer type [*SEM*]		
Experiment and measure	All single	All plural	All selected
Experiment 1			
Accuracy	0.41 (0.02)	0.78 (0.02)	0.66 (0.03)
Confidence	53.40 (3.25)	64.11 (3.43)	63.56 (3.32)
Experiment 2			
Accuracy	0.28 (0.02)	0.73 (0.03)	0.59 (0.04)
Rated likelihood	49.80 (3.55)	57.27 (3.97)	58.53 (3.87)

Note. Means for all single and all plural were computed from all answers of a given type, regardless of whether they were later selected for report. Means for selected answers were based on participants' selections for reporting, regardless of whether they were single or plural.

if the selected single answers had higher accuracy than the rejected single answers. This effect would show that participants tended to select the single answer when single-answer accuracy was high and opt for the plural answer only when single-answer accuracy was low. Indeed, as expected, this effect was significant, $t(23) = 5.08$, $p < .001$, $d = 1.37$. In contrast, the difference between selected and rejected plural answers was also significant, $t(23) = 2.67$, $p = .01$, $d = 0.79$, but the conditions were ordered in the opposite way than for single answers, giving rise to the interaction. Finally, we examined the accuracy for just the selected answers for reasons that will become evident later. Consistent with the main effect of answer type reported earlier, accuracy was higher for selected plural than for selected single answers, $t(23) = 2.32$, $p = .03$, $d = 0.64$.

Mean confidence. The bottom panel of Figure 1 shows mean confidence. We ran the same analyses as for accuracy to examine the role of confidence when selecting single versus plural answers. A 2 (answer type: single or plural) × 2 (decision: selected or rejected) repeated-measures ANOVA revealed significant main effects of answer type, $F(1, 23) = 36.10$, $MSE = 56.93$, $p < .001$, $h^2 = .61$, and decision, $F(1, 23) = 52.17$, $MSE = 29.64$, $p < .001$, $h^2 = .69$.

Confidence was higher for plural answers ($M = 67.50$, $SEM = 5.01$) than for single answers ($M = 58.25$, $SEM = 5.82$), showing that participants understood both the task and the basic law of probability that plural answers must be more likely to be accurate than single answers (although see below in this section). Confidence was also higher for selected answers ($M = 66.89$, $SEM = 4.83$) than for rejected answers ($M = 58.86$, $SEM = 6.00$). As with the accuracy analysis, the interaction was also significant, $F(1, 23) = 56.38$, $MSE = 648.00$, $p < .001$, $h^2 = .71$.

As previously, we conducted some simple follow-up tests to examine the interaction in more detail. If confidence guides or mediates the decision to select a single answer, then we would

expect higher confidence for selected single answers than for rejected single answers. Indeed, this was the case. Confidence for selected single answers was much higher than for rejected single answers, $t(23) = 9.07$, $p < .001$, $d = 2.96$. This result, as for accuracy, suggests that participants selected the single answer when confidence in it was high, but opted for the plural answer when confidence in the single answer was low. In contrast, the opposite ordering of conditions was observed for plural answers, $t(23) = 5.70$, $p < .001$, $d = 1.62$, giving rise to the interaction described earlier.

Finally, we examined the confidence for just the selected answers. In contrast to the results with accuracy, and qualifying participants' understanding of basic probability laws suggested by the main effect of answer type from the ANOVA, the confidence for selected single answers was higher than that for selected plural answers, $t(23) = 5.19$, $p < .001$, $d = 1.61$. This result, coupled with the accuracy results, shows that a dissociation exists between confidence in accuracy for selected answers. The increase in the number of options of the selected answer from one to three had the logical consequence that accuracy increased; however, counterintuitively, increased size for selected answers was associated with decreased confidence.

Calibration and resolution. To further examine the relation- ship between confidence and accuracy, we investigated calibration and resolution. *Calibration,* or *absolute monitoring accuracy,* is the degree to which rated confidence corresponds to actual accuracy and is often examined with *calibration curves*, a plot of accuracy (*y*-axis) at different levels of confidence (*x*-axis). In contrast, *resolution,* or *relative monitoring accuracy,* is the degree to which confidence discriminates between correct versus incorrect answers within participants.

Calibration curves for Experiment 1 are shown in the top panel of Figure 2. Initial analyses suggested that there were too few

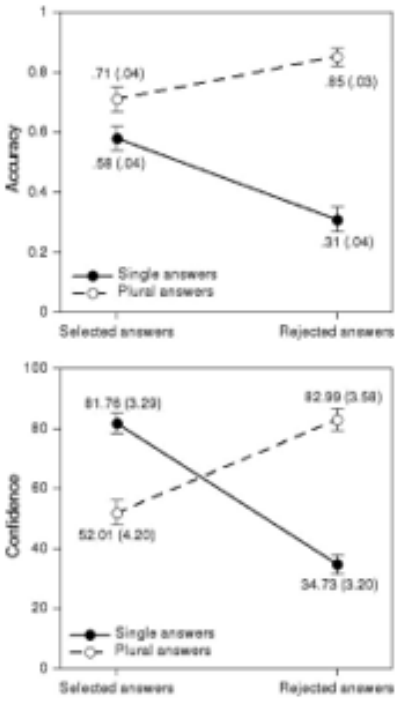

Figure 1. Mean accuracy, mean confidence, and standard errors of the mean (in parentheses) for selected and rejected single and plural answers in Experiment 1. For any given question, the selected answer is paired with the rejected answer of the opposite answer type.

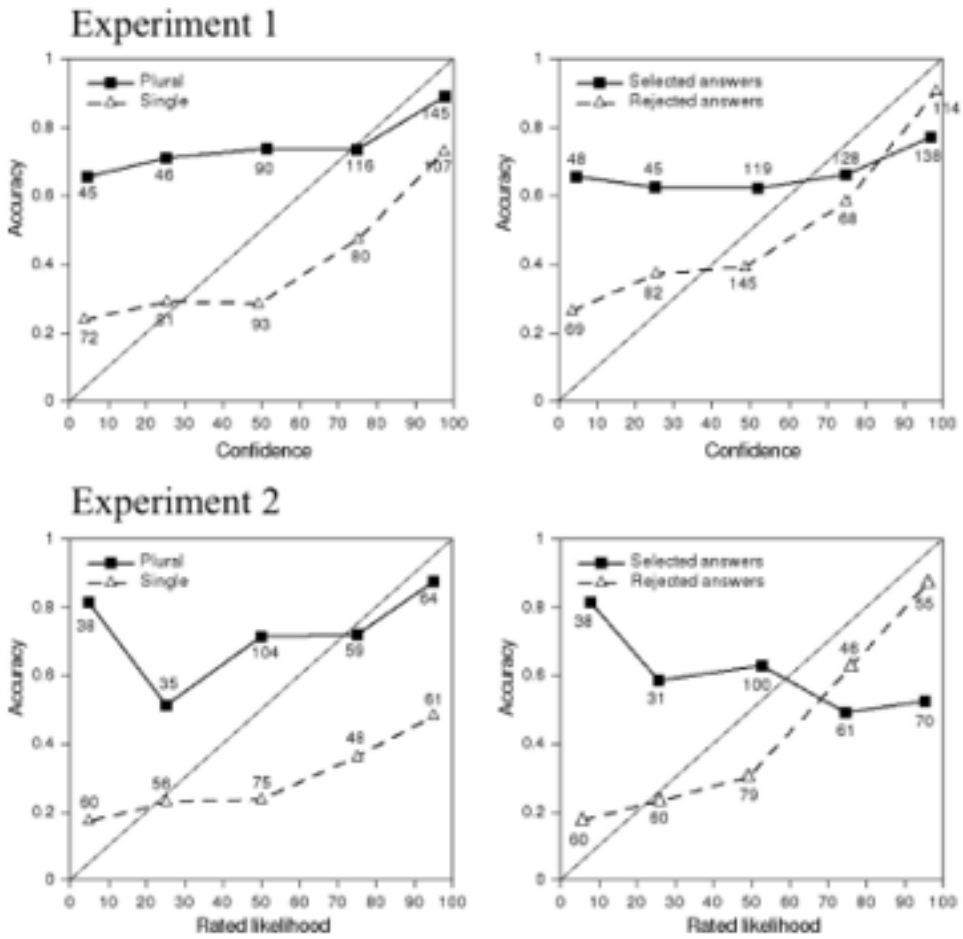

Figure 2. Calibration curves for Experiments 1 (top panel) and 2 (bottom panel). Within each experiment, the graph on the left shows calibration for all single and all plural answers, whereas the graph on the right shows calibration for all selected and all rejected answers. The number of observations in each category appears beside each data point. The diagonal line represents perfect calibration.

observations to plot separate curves for selected plural, rejected plural, selected single, and rejected single answers. Instead, we compared the answer types (collapsed across decision types) in the left-hand graph and the decision types (collapsed across answer types) in the right-hand graph. There were also too few observations in some of the confidence deciles to render a meaningful calibration curve, so we collapsed the 11 categories of confidence (0, 10 . . . 90, 100) into five categories: 0 –10, 20 –30, 40 – 60, 70 – 80, and 90 –100. The average confidence within each category was used as a marker for the category.

For the (left-hand) curves comparing answer types, there was mostly overconfidence for the single answers but underconfidence for the plural answers. This result is similar to those of previous studies (Goldsmith et al., 2002, 2005; Weber & Brewer, 2008), thus showing again that the plurality option and the procedure used here with qualitative answers yields comparable results to previous research on the grain-size option. For the (right-hand) curves comparing decision types, the results are consistent with the mean analysis presented earlier. That is, whereas rejected answers are reasonably well calibrated (i.e., close to the diagonal), selected answers are not. This poor calibration for selected answers is indicated by the fact that as confidence increased, accuracy remained constant. The near-zero slope of the curve for selected items probably reflects the fact that it consists of both single and plural answers. That is, for plural answers accuracy was high, but confidence was low, bringing the left-hand end of the curve upward, whereas the opposite was true of single answers, bringing the right-hand

end of the curve downward. Thus, coupled with the results shown in Figure 1, the calibration analyses indicate that the over- and underconfidence is limited to selected answers.

Finally, we calculated a resolution measure. The most common resolution measure in the metacognitive literature is the Goodman–Kruskal gamma correlation coefficient. However, gamma has recently been shown to have some very undesirable properties (see Masson & Rotello, 2009; Rotello, Masson, & Verde, 2008). As an alternative, Masson and Rotello (2009) recommended using an accuracy discrimination measure derived from receiver operating characteristic curves, which are used in signal detection analysis. To this end, we computed Type 2 receiver operating characteristics curves for the current data using the 11 confidence categories (see Higham, 2007, Appendix, for computational procedures). The area under the receiver operating characteristic curve (the AUC) is a "good index of sensitivity" (Macmillan & Creelman, 2005, p. 64), and it can easily be computed using the trapezoidal rule (e.g., Green & Moses, 1966; Pollack, Norman, & Galanter, 1964). Chance discrimination and perfect discrimination correspond to AUCs of 0.50 and 1.0, respectively.

Consistent with the calibration results, AUC was significantly higher for rejected answers (M = .73, SEM = .03) than for selected answers (M = .53, SEM = .03), $t(23) = 4.98$, $p < .001$, $d = 1.37$. Resolution for rejected answers was greater than chance (.50), $t(23) = 7.45$, $p < .001$, $d = 1.59$, whereas it was not for selected answers, $t(23) = 0.89$, $p = .39$, $d = 0.18$. In terms of a comparison of the answer types, resolution for all single answers (M = .66, SEM = .02) was greater than for all plural answers (M = .58, SEM = .03), $t(23) = 2.32$, $p = .03$, $d = 0.58$. Both

AUCs were above chance, $t(23) = 7.58$, $p < .001$, $d = 1.61$, for single answers and $t(23) = 2.45$, $p = .02$, $d = 0.52$, for plural answers. Overall, these results support the main conclusion that confidence was a particularly poor predictor of accuracy if an answer was selected.

Discussion. Two main conclusions can be drawn from this experiment. First, the analysis of accuracy with the plurality option replicated analogous studies that have examined the regulation of accuracy using the standard grain-size paradigm (e.g., Ackerman & Goldsmith, 2008; Weber & Brewer, 2008): Participants con- trolled the number of alternatives of their volunteered answer in an attempt to simultaneously maximize both accuracy and informativeness. This was shown in that when accuracy and confidence in the single answer were low, participants were more likely to select the more imprecise plural answer in an attempt to increase accuracy at the cost of informativeness.

The second result is that confidence was higher for selected single answers than for selected plural answers, despite the fact that accuracy was lower. Most of the research on the confidence– accuracy relationship has shown that it is positive or null depending on the domain. With general knowledge questions, the confidence–accuracy relationship is typically positive and strong (Costermans, Lories, & Ansay, 1992; Perfect, 2004). Strong positive relationships are also found in the eyewitness memory do- main, but small or null relationships are also observed depending on the conditions and variables manipulated (e.g., Brown, 2003; Higham, Luna, & Blank, 2011; Higham, Luna, & Bloomfield, in press; Odinot & Wolters, 2006; Perfect, 2004; Perfect & Hollins, 1996). However, our data showed a dissociation between confidence and accuracy for selected answers. This dissociation could be attributed to the combined effects of overconfidence for selected single answers and underconfidence for selected plural answers, although rejected answers showed almost no overconfidence or underconfidence.

These results with mean accuracy and confidence were sup- ported by the calibration and resolution analyses. Calibration was generally quite good for rejected answers, but very poor for selected answers. Similarly, the resolution analysis showed that discrimination between correct and incorrect answers was good for rejected answers, but no different from chance for selected answers. The net result was that although overall confidence was higher for all plural versus all single answers, participants drastically underestimated the effect of answer size on accuracy. If the plural answer was selected, the accuracy advantage of plural answers over their corresponding rejected single answers was 40%, whereas the confidence increase was only 17%. Conversely, if a single answer was selected, increasing answer size showed no advantage at all as far as participants were concerned. That is, as answer size increased, the accuracy advantage was 27%, whereas confidence only increased by 1% (see Figure 1).

Experiment 2

In the literature on grain-size control, there are slight but potentially critical differences in how confidence ratings are elicited. On one hand, Goldsmith and colleagues (Ackerman & Goldsmith, 2008; Goldsmith et al., 2002, 2005) required participants to rate the objective likelihood that a given answer is correct. For example, Goldsmith et al. (2002, Experiment 2) asked participants to "estimate the chances that the answer contains the true value" (p. 78). On the other hand, other grain-size control researchers have asked participants to rate their confidence rather than the chances of accuracy. For example, Weber and Brewer (2008, Experiment 1) required participants to indicate "their confidence that their answer contained the true value" (p. 53). In both cases, ratings were made on a 100-point (percentage) scale. This latter instruction is similar to the one we used in Experiment 1.

Although the two methods of eliciting confidence ratings are superficially very similar, the chance of an important difference remains. For example, when asked to rate "confidence," participants may actually have rated their subjective knowing of the (single) answer rather than the objective likelihood of answer correctness. Thus, with respect to the plurality option, in Experiment 1 participants may have reasoned something similar to "I know that my three-alternative (plural) answer is likely to be objectively correct, but I'm going to rate my confidence low because I don't really feel like I know the (single) answer to the question." If so, this reasoning could have produced the substantial underconfidence observed for selected plural answers and would partially explain the confidence–accuracy dissociation that was observed.

To address this potentially important distinction between rated confidence and rated likelihood, in Experiment 2 we replicated Experiment 1 with two important changes. First, to reduce the possibility that confidence ratings actually pertained to how much participants believed they knew the single answers to questions rather than the objective probability that the answer was correct, the instructions were altered. In particular, references to *confidence* were replaced by *likelihood.* We reasoned that the term *likelihood* would focus participants on the objective reality of the situation rather than the subjective status of their knowledge. Thus, for plural answers, participants were instructed to "indicate the likelihood that the correct answer is any one of the three alternatives." To further emphasize the objective nature of the rating, participants were reminded that answers with more alternatives were more likely to be correct

The second important change was the use of different experi- mental materials. In particular, we adapted the slideshow of a simulated murder scene used in Higham et al. (in press) for use in this experiment. The use of different materials allowed us to test whether the results we obtained with the multiple-choice method- ology were specific to the particular crime scene and test questions that were used in Experiment 1 or whether they were more gen- eralizable.

Method

Participants. Twenty participants (three men, 17 women, mean age = 19.05 years, *SD* = 0.76, range = 18 –21) from the University of Minho (Braga, Portugal) took part in this experiment in exchange for course credits.

Materials and procedure. We used a longer version of Higham et al.'s (in press) slideshow: Eighteen slides showed a car leaving a house, several rooms of the house, and finally the bathroom, with a large knife covered in blood in the sink and a dead body in the bathtub. Fifteen questions with five plausible alternative answers were developed. Participants were explicitly told that only one of the five alternatives for each question was correct. Questions addressed different details such as objects in the several rooms, a picture on a calendar, the shape of a mirror, and some characteristics of the knife.

The procedure was the same as for Experiment 1, with the aforementioned change in the instructions: Instead of asking par- ticipants to rate confidence, participants were asked to rate the likelihood that the answer selected was correct. Specifically, below the text box for the single answer, the instruction that appeared on the computer screen was to "indicate the likelihood that this answer is correct. 0% indicates that there is no likelihood at all, and 100% indicates that there is certainty that the option you chose is the correct answer." Similarly, below the three text boxes for the plural answer, participants were instructed to "indicate the likeli- hood that the correct answer is *any one of the three alternatives*. 0% indicates that there is no likelihood at all, and 100% indicates that there is certainty that the correct answer is one of the three you chose." Furthermore, the initial instructions also included a sen- tence to remind participants of the effect of base rates: "Keep in mind that you are more likely to select the correct alternative as you select more alternatives."

Results

Participants selected single and plural answers, respectively, 34% and 66% of the time. The proportion of single answers that was selected was not significantly different from that found in Experiment 1, $t(42) = 1.53$, $p = .13$, $d = 0.47$. Descriptive statistics for accuracy and confidence for all the single answers, all the plural answers, and all the selected answers can be seen in Table 1.

Mean accuracy. Mean accuracy is shown in the top panel of Figure 3. A 2 (answer type: single or plural) × 2 (decision: selected or rejected) repeated-measures ANOVA indicated that accuracy was higher for plural answers ($M = .75$, $SEM = .04$) than for single answers ($M = .33$, $SEM = .06$), $F(1, 19) = 174.34$, $MSE = 0.02$, $p < .001$, $h^2 = .90$. The main effect of decision was not significant; $F < 1$. However, as in Experiment 1, the

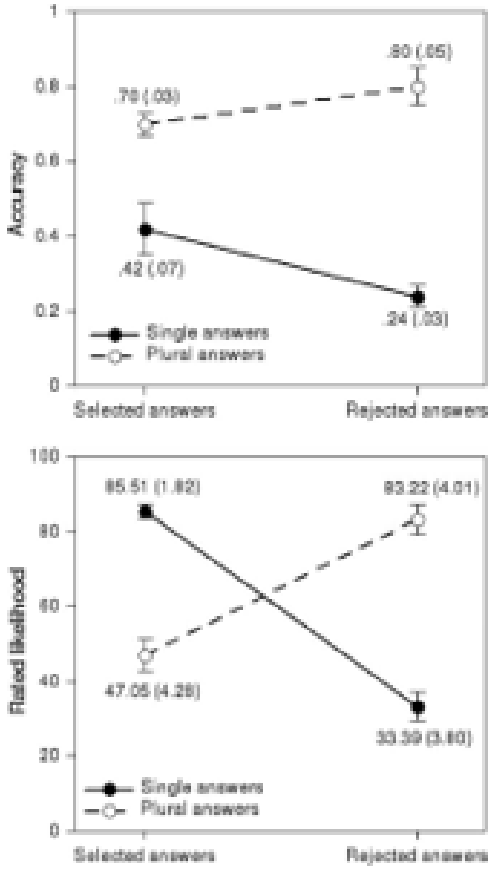

Figure 3. Mean accuracy, mean rated likelihood, and standard errors of the mean (in parentheses) for selected and rejected single and plural answers in Experiment 2. For any given question, the selected answer is paired with the rejected answer of the opposite answer type.

interaction was significant, $F(1, 19) = 6.44$, $MSE = 0.06$, $p = .020$, $h^2 = .25$.

To investigate the interaction further, we conducted some follow-up tests. Accuracy was higher for the selected single answers than for the rejected single answers, $t(19) = 2.16$, $p = .04$, $d = 0.70$, whereas the reverse was true of plural answers, $t(19) = 2.15$, $p = .04$, $d = 0.57$, giving rise to the interaction. Following the same reasoning as in Experiment 1, the former effect indicated that participants successfully regulated memory accuracy with the plurality option. As in Experiment 1, plural answers had a significant accuracy advantage over single answers when only selected answers were analyzed, $t(19) = 3.80$, $p = .001$, $d = 1.08$.

Mean rated likelihood. Mean rated likelihood is shown in the bottom panel of Figure 3. A 2 (answer type: single or plural) × 2 (decision: selected or rejected) repeated-measures ANOVA revealed significant main effects of answer type, $F(1, 19) = 6.02$, $MSE = 107.29$, $p = .02$, $h^2 = .24$, and decision, $F(1, 19) = 18.47$, $MSE = 68.9$, $p < .001$, $h^2 = .49$. Rated likelihood was higher for plural answers ($M = 65.14$, $SEM = 5.79$) than for single answers ($M = 59.45$, $SEM = 6.59$) and also higher for selected answers ($M = 66.28$, $SEM = 5.43$) than for rejected answers ($M = 58.30$, $SEM = 6.83$). The interaction was also significant, $F(1, 19) = 117.55$, $MSE = 332.00$, $p < .001$, $h^2 = .86$.

We conducted follow-up tests to investigate the interaction. As with accuracy, rated likelihood was higher for selected single answers than for rejected single answers, $t(19) = 13.72$, $p < .001$,

d = 3.91, whereas the reverse was true for plural answers, *t*(19) = 7.15, *p* < .001, *d* = 1.95, giving rise to the interaction.

In contrast to accuracy, rated likelihood was higher for selected single answers than for selected plural answers, *t*(19) = 10.24, *p* < .001, *d* = 2.62. This result, coupled with the results found for accuracy, indicates that this experiment replicated the confidence– accuracy dissociation for selected answers found in Experiment 1. This replication occurred despite the fact that participants were asked to rate the objective probability of accuracy rather than subjective confidence.

Calibration and resolution. As in Experiment 1, to investigate the relationship between accuracy and confidence in more detail, we examined both calibration and resolution. Calibration curves analogous to those computed in Experiment 1 are shown in the bottom panel of Figure 2. The (left-hand) curves comparing answer types showed the same pattern of overconfidence for single answers and underconfidence for plural answers that we observed in Experiment 1. Also similar to Experiment 1, the (right-hand) curves comparing decision types indicated that calibration was reasonably good for rejected answers, but appalling for selected answers. Indeed, calibration was so poor for these answers that accuracy tended to decrease as rated likelihood increased. This decrease differs from the analogous calibration curve in Experiment 1, which was flat. This negative slope for selected answers in this experiment probably occurred because single-answer accuracy was somewhat lower in this experiment than in Experiment 1. Because these items were associated with high confidence, whereas plural answers were associated with lower confidence, the right-hand part of the curve would have been brought downward. Overall, the calibration results are consistent with the dissociation reported for the analysis of mean accuracy and confidence.

For the resolution analysis, AUC for rejected answers was reasonably good (*M* = .73, *SEM* = .03) and much better than for selected answers (*M* = .39, *SEM* = .03), *t*(19) = 7.47, *p* < .001, *d* = 2.46. Indeed, whereas resolution for rejected answers was greater than chance (.50), *t*(19) = 8.26, *p* < .001, *d* = 1.94, it was below chance for selected answers, *t*(19) = 3.09, *p* = .003, *d* = 0.73. These results add support for the negative relationship re- ported earlier between accuracy and confidence for selected answers. In terms of comparing answer types, resolution for all single answers (*M* = .59, *SEM* = .04) did not differ from that for all plural answers (*M* = .54, *SEM* = .04), *t*(18) = 0.79, *p* = .44, *d* = 0.30. Resolution for single answers was greater than chance, *t*(18) = 2.33, *p* = .03, *d* = 0.56, whereas that for plural answers was not, *t*(18) = 1.09, *p* = .29, *d* = 0.26.[2] The poor resolution for the different answer types probably results from the fact that each type consisted of both rejected and selected answers, which have opposing resolution scores (one above chance; the other below chance). Thus, when they are mixed together, the scores may have cancelled each other out to yield chance performance.

Discussion

Experiment 2 replicated the main results obtained in Experiment not limited to the particular materials used in that experiment. Overall, both rated likelihood and accuracy were higher for plural answers than for single answers, and participants used these differences to strategically regulate accuracy.

More important, the results were obtained in this experiment despite changes to the instructions to ensure that when rating confidence, participants were rating the objective

likelihood of given answers' accuracy rather than participants' subjective sense of knowing. Indeed, if anything, the confidence–accuracy dissociation was even more pronounced in this experiment than in Experiment 1, demonstrated by below-chance resolution and a calibration curve with a negative slope for selected answers. The comparable results between our experiments, as well as the com- parable results obtained between researchers who use confidence rating instructions (e.g., Weber & Brewer, 2008) versus instructions to rate the chances of accuracy (e.g., Goldsmith et al., 2002), suggest that the distinction is not important as far as participants are concerned.

General Discussion

In two experiments, we investigated the plurality option, an accuracy regulation mechanism incorporating strategic control over the number of alternatives contained in an answer. We have argued that the plurality option should be considered a specific instantiation of the broader process of grain-size control. In sup- port of this viewpoint, we obtained comparable results between our experiments on plurality and previous research on grain-size control despite substantial methodological differences. The comparable results also suggest that people are quite versatile in terms of making use of any option available to them in regulating the accuracy of their answers.

Some readers, however, may argue that plurality control is sufficiently different from previous conceptualizations of grain- size control to be considered a new regulatory mechanism in its own right. For example, previous research on grain-size control has relied exclusively on self-generated answers to quantitative questions involving numeric answers (e.g., age, time, distance, height; Goldsmith et al., 2002) or gradients (e.g., dark color vs. navy blue; Weber & Brewer, 2008), whereas ours did not. Whether one considers the plurality option to be a stand-alone mechanism or not, the important point is that our research on plurality is the first to demonstrate that people can regulate accuracy by varying the number of qualitatively different, even contradictory, alternatives in their answers. Unlike previous investigations on grain-size control, the plurality option is potentially highly adaptable, which is beneficial to both rememberers, who are able to regulate the accuracy of their answers in a variety of different contexts, and to researchers, who can investigate the strategic regulation of accuracy using a host of materials and paradigms.

Given the versatility of the plurality option, eyewitnesses could potentially rely on it to maintain good memory accuracy in a variety of different scenarios in which attention or memory has been experimentally compromised. Such examples include compensating for overly restrictive encoding due to weapon focus (Loftus, Loftus, & Messo, 1987; Steblay, 1992) or correcting source-monitoring errors due to receiving postevent misinformation (e.g., Higham, 1998; Luna & Migueles, 2008, 2009). By increasing the number of alternatives reported, the negative memory effects observed in these paradigms could possibly be regulated, although at a cost of informativeness. Source-monitoring errors like those seen in misinformation paradigms may lend themselves particularly well to being "saved" by the plurality option. For example, eyewitnesses may confidently remember that both a red jacket and a green jacket were encountered before a memory test, but they may not specifically remember the color of the robber's jacket as shown in the original event. If faced with both the red- and the green-answer alternatives on a multiple- choice memory test, the error rate may be high if a single answer must be chosen. However, if both the red and green alternatives

can be included with a third one to create a plural answer, as in our experimental methodology, source-monitoring errors may be less pronounced, or even eliminated altogether. Note that such error correction is not possible with the types of questions that have typified previous research on single-answer grain-size control.

Several other instantiations of grain-size control may potentially be worth investigating in future research. For example, Higham et al. (in press, Experiment 2) investigated performance on a multiple-choice test about a witnessed event when single- alternative answers could be provided at either a superordinate level of a category (e.g., a single-story building) or at a more basic level of a category (e.g., a bungalow). They found that accuracy incentives affected the degree to which each answer type was chosen. That is, high accuracy incentives led to more superordinate-level (coarse-grained) responses than low incentives, indicating that the level of the category at which participants answered questions was under strategic control. However, because the different category levels of the alternatives were used to answer different research questions than those asked in the current work, Higham et al.'s data cannot be used to test whether participants were able to use the category option to strategically regulate accuracy. It may well be the case that people readily control the category level of their answers to regulate accuracy in the same way that they control quantifiable boundaries around answers and the number of alternatives in their answers.

Limitations on the Use of the Plurality Option

In these experiments, participants were presented with several candidate answers that they selected to build their single and plural answers. This procedure may be restrictive in one sense because, for some questions, participants may not believe the correct answer is among the alternatives. In this vein, an interesting avenue for future research may be to maintain the multiple-choice testing methodology but to include options such as "none of the above" or "I don't know" (report option). However, nothing limits future plurality option research to the multiple-choice testing format. It is conceivable that a recall version of the task could be developed by requiring participants to self-generate single and plural answers. Another interesting variation of the plurality option procedure would be to present several incorrect alternatives or to tell participants that the correct answer may or may not be present as an alternative. This way, the procedure may more closely resemble an actual eyewitness interrogation in which there is greater uncertainly about answer accuracy; that is, neither interviewer nor interviewee necessarily knows the correct answer to the questions posed by the interviewer.

Previous research has shown that although participants can regulate accuracy by controlling the grain size of their answers, that regulation is far from optimal (e.g., Goldsmith et al., 2002, 2005; Weber & Brewer, 2008). Our results with the plurality option concur with this finding. First, for Experiments 1 and 2, 27% and 38%, respectively, of the selected-but-incorrect single answers were correct at the plural level. These data show that about one third of time that participants selected an incorrect single answer, the loss of accuracy could have been avoided by selecting the plural answer. Similarly, for Experiments 1 and 2, 31% and 24%, respectively, of the plural answers that were selected and correct would also have been correct if the single answer was selected. These data suggest that participants sometimes suffered the cost of low informativeness when it was not necessary; that is, they selected the plural answer even though the single answer was correct.[3] Both results suggest metacognitive monitoring that is

far from perfect, which limits the extent to which accuracy can be regulated (Koriat & Goldsmith, 1996).

Second, although our participants demonstrated sensitivity to the beneficial effect that augmenting the number of alternatives would have on accuracy—indeed, they would not have been able to regulate accuracy with the plurality option without some of this sensitivity—subjective confidence indicated that they were not as sensitive as they should have been. As a case in point, whereas the difference in accuracy between all plural and all single answers in Experiment 2 was 45%, the analogous difference in confidence was only 7%, more than a sixfold difference. The fact that confidence so profoundly underestimated the impact of the number of alternatives on accuracy means that participants are miscalibrated and cannot fully benefit from the regulatory effects of the plurality option.

An additional challenge for the eyewitness would be to identify cases in which memory impairment is likely, suggesting that adding alternatives to the answer could be potentially beneficial, compared with cases in which a single answer is acceptable. In our own work on the misinformation effect (e.g., Higham et al., 2011), we found that people are unable or unwilling to change regulatory strategies on an item-by-item basis even if they are told which questions pertain to information about which they have been misinformed and which questions do not. Taken together, all these findings suggest that, although participants can in principle limit the deleterious effects of myriad factors on memory by controlling factors such as report criterion, grain size, and answer size, how this control can be applied has serious limitations.

Relationship Between Confidence and Accuracy

In both experiments, accuracy for selected plural answers showed a marked and clear advantage over selected single answers. However, the confidence ratings showed the opposite ordering, that is, confidence was higher for selected single than for selected plural answers, producing a dramatic confidence– accuracy dissociation. Confidence–accuracy dissociations of the sort observed here are not common in the study of memory although they are not unprecedented. For instance, in an experiment on recognition memory, Busey, Tunnicliff, Loftus, and Loftus (2000, Experiment 3) presented faces in a study phase under dim or bright luminance conditions and the same faces again under either dim or bright luminance at test. Recognition for faces presented dim and tested bright was lower, but confidence was higher than faces presented and tested under low-luminance conditions. Participants assumed that higher luminance helped with recognition, and as a consequence, they rated the bright-luminance conditions with higher confidence, when actually the effect on accuracy for those conditions was the opposite. Similarly, in a meta-analysis of 14 experiments, Chandler (1994) found that showing a picture related to a target during the study phase decreased or had no effect on later recognition, but it did increase confidence.

The confidence–accuracy dissociation we observed in our re- search was reflected in the overconfidence on the selected single answers and the underconfidence on the selected plural answers.[4] Potentially, both forms of bias can be accounted for by considering participants' metacognitive beliefs about their personal knowledge states. In particular, we hypothesize that personal theories of knowledge tend to be overly dichotomous, failing to incorporate the continuous, graded form that knowledge often assumes (as indexed by accuracy). For the current situation, people's dichotomous metacognitive theories may falsely lead them to believe that they either definitively know the answer to a question, or they have no knowledge

whatsoever. If participants conclude that they know the answer, they select a single answer and rate confidence overly high. If they conclude that knowledge is lacking, they select a plural answer and rate it overly likely to be wrong. In other words, their ratings tend to be anchored (Tversky & Kahneman, 1974) at 100% confidence for "known" single answers and 0% confidence for "not known" plural answers. Ratings may be adjusted slightly from those anchors according to other factors such as the (un)familiarity of the alternatives or the amount of domain knowledge, but the net result is the pattern of overconfidence for selected singular answers coupled with underconfidence for selected plural answers that we observed. Note that unbiased calibration for rejected answers possibly stems from the fact that participants have not assessed these answers with respect to their subjective knowledge state as they have with selected answers.

In support of the hypothesis that the knowledge states are dichotomous rather than continuous, Higham (2007) found that participants tended to be largely insensitive to penalty variations when answering aptitude questions and given a report option. Such a result might occur if participants reasoned that an answer to a question is either known, in which case the likelihood of a correct answer is 100%, or not known, in which case there is only chance probability of accuracy. If so, and students also assumed that these states are well monitored, then there would be no reason for them to respond differently because the penalty was varied: Their best strategy with either point system would be to simply report the known answers and withhold the not-known answers.

Conclusions

People can control the grain size of their memory reports to maintain reasonable levels of output accuracy even though memory retention is variable. In this article, we have examined a new instantiation of the grain-size regulatory mechanism that we have dubbed the *plurality option.* The procedure used to study this option is adaptable and can be used with many different types of questions, opening up new ways of investigating accuracy regulation in different contexts. The application of the procedure to an eyewitness scenario in the current experiments has shown that although participants recognize that plural answers are more likely to be correct than single ones and that they can use this fact to regulate memory accuracy, the real impact of adding alternatives to an answer is only fully appreciated if participants feel they do not know the single answer. However, our results have shown that increased answer size is beneficial to accuracy regardless of confidence in the single answer. Future research might focus on teaching participants about the true impact of answer size on accuracy so that it can be more effectively used to regulate accuracy.

---

[4] Critics may argue that the substantial overconfidence that we observed for selected plural answers in both experiments can be partially attributed to the fact that we presented the full 0%–100% confidence–likelihood scale to participants when they were asked to create answers. Instead, critics may argue, the bottom end of the scale should start at the chance likelihood for that answer size (i.e., 20% and 60%, respectively, for single and plural answers). By allowing participants to make ratings less than chance level, we may have misled them in that such ratings logically mean that other, nonchosen alternatives should have been included in the answer. However, in our view, it is not the experimenter's job to "force" logical responding

from participants by doctoring the scale on which they are responding. If it is, then one needs to consider removal of other scale values as well. For example, the large body of evidence demonstrating the fallibility of memory suggests that participants should logically never be allowed to indicate 100% certainty in their memory accuracy. There can be no doubt that had we presented participants with a 20%–100% scale for single answers and a 60%–100% scale for plural answers, the pattern of overconfidence and underconfidence that we observed would be altered. In all likelihood, overconfidence would have decreased and overconfidence would have increased. However, it is not clear how such changes to overconfidence and underconfidence should be interpreted, particularly given that scales for plural versus single answers would no longer be directly comparable

References

Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting—with and without satisficing knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1224 –1245.

Ben-Simon, A., Budescu, D., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21,* 65– 88.

Brown, J. M. (2003). Eyewitness memory for arousing events: Putting things into context. *Applied Cognitive Psychology, 17,* 93–106.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psycho- nomic Bulletin & Review, 7,* 26 – 48.

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

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.

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2,* 79 –96.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory.* Mahwah, NJ: Erlbaum.

Masson, M., & Rotello, C. (2009). Sources of bias in the Goodman– Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 509 –527.

Odinot, G., & Wolters, G. (2006). Repeated recall, retention interval and the accuracy-confidence relation in eyewitness memory. *Applied Cognitive Psychology, 20,* 973–985.

Perfect, T. J. (2004). The role of self-rated ability in the accuracy of confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology, 18,* 157–168.

Perfect, T. J., & Hollins, T. S. (1996). Predictive feeling of knowing judgements and postdictive confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology, 10,* 371–382.

Pollack, I., Norman, D., & Galanter, E. (1964). An efficient nonparametric analysis of recognition memory. *Psychonomic Science, 1,* 327–328.

Rotello, C., Masson, M., & Verde, M. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psycho- physics, 70,* 389 – 401.

Steblay, N. M. (1992). A meta-analytic review of the weapon focus effect.

Law and Human Behavior, 16, 413– 424.

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

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

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