Using the regulation of accuracy to study performance when the correct answer is not known

KARLOS LUNA[1] and BEATRIZ MART´IN-LUENGO[2]

[1]*School of Psychology, University of Minho, Braga, Portugal*

[2]*National Research University-Higher School of Economics, Moscow, Russian Federation*

Luna, K. & Martín-Luengo, B. (2017). Using the regulation of accuracy to study performance when the correct answer is not known. *Scandinavian Journal of Psychology, 58,* 275–283.

We examined memory performance in multiple-choice questions when correct answers were not always present. How do participants answer when they are aware that the correct alternative may not be present? To answer this question we allowed participants to decide on the number of alternatives in their final answer (the plurality option), and whether they wanted to report or withhold their answer (report option). We also studied the memory benefits when both the plurality and the report options were available. In two experiments participants watched a crime and then answered questions with five alternatives. Half of the questions were presented with the correct alternative and half were not. Participants selected one alternative and rated confidence, then selected three alternatives and again rated confidence, and finally indicated whether they preferred the answer with one or with three alternatives (plurality option). Lastly, they decided whether to report or withhold the answer (report option). Results showed that participants' confidence in their selections was higher, that they chose more single answers, and that they preferred to report more often when the correct alternative was presented. We also attempted to classify *a posteriori* questions as either presented with or without the correct alternative from participants' selection. Classification was better than chance, and encouraging, but the forensic application of the classification technique is still limited since there was a large percentage of responses that were incorrectly classified. Our results also showed that the memory benefits of both plurality and report options overlap.

Key words: Metamemory, eyewitness memory, plurality option, report option.

Karlos Luna, *Psychology Research Centre*, *School of Psychology*, *University of Minho*, *Campus de Gualtar*, *4710-057*, *Braga*, *Portugal*. *Tel.: +351 253 604 220; fax: +351 253 604 224;* e-mail: karlos.luna.ortega@gmail.com.

INTRODUCTION

Much research in psychology has been devoted to how decisions are made when the outcome is not certain (e.g., there is 50% chance of getting a reward). In memory studies, uncertainty about the outcome means that the respondent of a question usually does not know whether the response is correct or not (e.g., in an exam). In this research, we added another layer of uncertainty and studied memory performance for recognition questions for which the correct alternative was not always presented. In real life this may occur, for example, when police do not know what happened and question a witness. In the present research, our main objective was to study the metamemory processes when participants are aware that the correct answer may not be presented in a multiple-choice question.

In this research, we presented multiple-choice questions that

sometimes did not include the correct alternative. Other studies have used procedures also related to this topic. For example, in a study focused on the effect of test order, Hollins and Weber (2017) asked participants multiple-choice general knowledge questions and presented the correct alternative for only half of them. Other researchers have asked unanswerable questions, for example, questions that cannot be answered from the information presented (Pezdek, Lam, & Sperry, 2009; Pezdek, Sperry, & Owen, 2007; Scoboria, Mazzoni, & Kirsch, 2008; Waterman, Blades, & Spencer, 2001). Most of these studies were aimed at analyzing speculation and confabulation when there is no memory, and none of them tried to represent that the questioner sometimes does not know the answer or does not want to provide it Similarly, studies on witness identification often include a target-absent condition, in which the perpetrator is not present (target-absent lineups), to control for response bias (Wells & Turtle, 1986). In identification studies, metamemory, and in particular the confidence-accuracy relationship, has been an important focus of research. However, analyses are usually reported with target-present and target-absent lineup conditions collapsed (e.g., Lindsay, Read, & Sharma, 1998) or analyzed separately (e.g., Brewer, Keast, & Rishworth, 2002; Brewer & Wells, 2006). To our knowledge, there are no studies that compare metamemory performance for target-absent and target- present conditions. Thus, little can be learnt from studies on confabulation or witness identification about the effect of the presentation of the correct alternative on metamemory and on the processes leading to a response.

To study the control processes under uncertainty about the presence of the correct alternative, we asked participants to answer multiple-choice questions for which the correct answer was not always present. The reader may wonder how memory can be studied if there is no way to select the correct answer. Without the correct answer it is not possible to compute any measure of accuracy. Accuracy is arguably the most important measure in studies on memory, but it is not the only one. Other measures are also very informative about memory processes, such as frequencies or confidence. To study participants' performance when there is no correct response we resorted to an area in which frequencies and confidence play a major role: the regulation of accuracy (Ackerman & Goldsmith, 2008; Koriat & Goldsmith, 1996; Higham, 2007; Horry, Brewer, & Weber, 2016; Luna, Higham, & Mart´ın-Luengo, 2011). The regulation of accuracy is a process, based on the monitoring and control of our memories that allows accuracy to be maximized by varying several dimensions of the response provided.

The regulation of accuracy can be studied through a set of techniques that allow respondents to increase or decrease accuracy by filtering out answers with low chances of being correct (Higham, 2013; Luna, Mart´ın-Luengo, & Brewer, 2015; for a review, see Goldsmith, 2016). Two of these techniques are used here: the report option (Koriat & Goldsmith, 1996), and the plurality option (Luna *et al*., 2011). For example, after a bank robbery police may ask about how the robber concealed his face, and offer five typical alternatives: a mask, a scarf, a stocking, a bandana, and a balaclava. In a typical event memory experiment in which the correct alternative is presented and the answer is forced, participants have to select one alternative. When the report option is allowed, participants have to decide whether they want to report or withhold their selection, that is, leave the question unanswered. In the plurality option participants are typically given the chance to select one answer (e.g., it was a mask; the single answer), or three answers (e.g., it was a mask, a scarf, or a balaclava; the plural answer), and then select which one, single or plural, they prefer to report.

Broadly speaking, the results of both the report and the plurality option are similar. Participants do not always select the same type of answer, sometimes choosing the single answer, or the option to report, and sometimes the plural answer, or the option to withhold (Higham, 2013; Higham, Luna, & Bloomfield, 2011; Luna & Mart´ın-Luengo, 2012b, 2017), and by doing so, the accuracy of the final report increases. Thus, even though the regulation of accuracy is mostly related to accuracy, frequency of selection of report/withhold and single/plural is key. In the same vein, both regulatory strategies are based on confidence ratings. Confidence has been shown to be one of the main factors that drive the selection between report/withhold (Koriat & Goldsmith, 1996) and single/plural (Luna *et al*., 2011). If the single answer is rated with low chances of being correct (or with low confidence, see Luna *et al*., 2011), then it is rejected (report option) or the plural answer is selected (plurality option). If the single answer is rated with high confidence, then it is reported. In sum, while the accuracy of the final answer is the most relevant outcome in the study of the regulation of accuracy, the frequency of selection of the different types of answers (report/withhold or single/plural) and the confidence ratings provide extremely valuable information.

Our main objective was to study performance when the correct answer was not presented. Without the correct answer, none of the alternatives presented will have a match in participants' memory, so we expected them to rate their selections with lower confidence than when the correct answer was presented. As the selection of single/ plural and report/withhold is heavily, but not exclusively, driven by confidence ratings (Koriat & Goldsmith, 1996; Luna *et al*., 2011, 2015), we also expected there to be fewer single and fewer reported answers for questions without a correct alternative than when the correct answer was presented. In partial support of our hypotheses, Hollins and Weber (2017), and Pezdek *et al*. (2007, 2009) found a higher rate of withhold answers (i.e., lower proportion of reported answers) for questions presented without a correct answer.

If our hypotheses are confirmed, a potentially relevant applied consequence is that it may be possible to classify a *posteriori questions* presented with and without a correct alternative. Posteriori classifications have been made in the past, using factors such as confidence, response latency, or feedback, mostly focused on classifying a witness identification as correct or incorrect (Sauer, Brewer, & Weber, 2008; Saulerland, Sagana, & Sporer, 2012; Smith, Lindsay, & Prike, 2000; Smith, Lindsay, Prike, & Dysart, 2001). For example, Smith et al. (2000) correctly classified 75% of the positive identifications. To our knowledge, no attempt at classification has been made for the memory of a full event (what Hollins & Perfect, 1997, called event memory). Thus, this research constitutes the first attempt to classify a question about a full criminal event as being presented with or without the correct response. We hypothesized that single answers: (1) were selected in the plurality option; (2) were assigned a confidence rating of 100; and (3) the participant decided to report (henceforth, S100R answers) occur more often for questions with a correct alternative than without. If this is confirmed, then it may be of interest to know how well S100R answers classify questions with and without the correct alternative. If classification is good, then questions with and without the correct response can be identified from the participant's selection. Importantly, it is not necessary to know the correct answer for this classification to work, and in a police investigation not knowing is the usual situation.

Another objective of this research was to explore the memory benefits when both options, plurality and report options, are allowed. The memory benefit (performance consequences in the terminology of Goldsmith, 2016) is defined, for the plurality option, as the accuracy

increase from all single answers to the accuracy of the final answer. For the report option, the memory benefit is the accuracy increase from the final answer to the reported answers. As it is only possible to compute memory benefits for questions presented with the correct alternative, our predictions are limited to these questions.

To our knowledge, no research has explored the memory benefit of both options together. Two experiments offered participants the chance to answer "don't know" in a regulation of accuracy experiment. Ackerman and Goldsmith (2008, Experiment 3) presented questions with numerical answers and asked participants to provide an interval of their preferred length, or to answer "don't know." Weber and Brewer (2008, Experiment 2) asked participants to produce single and plural answers, and then to choose between the single, the plural, and a "don't know" option. In both cases the memory benefit of both options together could not be computed.

Both the report and the plurality options are based on the same principles that drive the regulation of accuracy, for example, informativeness and confidence (Ackerman & Goldsmith, 2008; Luna et al., 2015). It may be that the memory benefit obtained by both options overlaps, and that when both are available there is no substantial increase in memory over what would be obtained if only one was used. Alternatively, despite similarities, the memory benefit of both options together may be higher than when separated. Luna and Mart´ın-Luengo (2017) suggested that the memory benefit of the plurality option comes from two sources: the ability to filter out answers with low chances of being correct, and the addition of a subset of answers with higher chances of being correct (namely, the plural answers). Only the first source is available in the report option. In the experiment reported here participants first answered the plurality option and then the report option (see the Procedure section). The report option first and then the plurality option alternative did not make much sense because if the answer is withheld in the report option, then there will not be much incentive to keep answering during the plurality option. Since the plurality option was presented first and has two sources for the memory benefit, we expected that most of the accuracy increase would be obtained with the plurality option and no substantial increase in accuracy would be obtained with the report option.

To meet our objectives and test our hypotheses, we conducted two experiments in which participants watched a crime and answered multiple-choice questions. Half of the questions were presented with the correct alternative and half without, allowing us to create the two main conditions. For each question, the participants provided several answers following the standard plurality option and then the report option. This research has potential applications in forensic situations, but the experiments were not designed with a direct application in mind. Our experimental choices were intended to increase experimental control and allow comparison with previous studies on the regulation of accuracy. Therefore, the procedure employed is somehow removed from a forensic situation (see below). We will come back to the generalization of this research to forensic scenarios in the General Discussion.

EXPERIMENT 1: METHOD

Participants and design

This experiment was conducted at a Turkish university and in Turkish. A total of 68 participants completed the experiment in exchange for course credits (61 females, mean age 20.10 years,

SD = 1.52). Participants were randomly allocated to the two counterbalanced conditions. The experiment was a two correct alternative (with, without) within-subject design.

Materials

A 3-minute excerpt from the film *The Stick-Up* (Herrington, 2002) was used. The video showed two security guards who take some sacks of money to the safe room of a bank and drive away. Afterwards, a robber walks into the bank, threatens customers and tellers, forces the branch director to fill a gym bag with the money, and drives away.

A subset of 28 questions with five alternatives from Luna *et al*. (2015, Experiment 2) were selected for the experiment. Both authors checked the questions to avoid misleading and deceptive ones. One of the alternatives was correct and the other four were incorrect but plausible. To create the experimental manipulation one new incorrect alternative per question was created. For questions presented with correct alternative, the correct and four incorrect alternatives were presented (e.g., What kind of gloves did the robber wear? Alternatives: latex gloves [correct response], leather gloves, woollen gloves, rubber gloves, or work gloves). For questions without the correct alternative the five alternatives were incorrect (e.g., latex gloves was replaced with cotton gloves).

Procedure

The participants completed the experiment on a computer and in groups of up to four. After giving consent and reporting basic demographics, they watched the video and completed a filler task for five minutes (a sudoku). Before the video, the participants were instructed to pay attention because they would be asked to answer questions about the event. This instruction was intended to match their expectations about the experiment. The video was presented without audio to avoid differential memorability or attention paid to some scenes because of the background music and accompanying sounds. The event was easy to follow without audio.

After the filler task, the participants read the instructions for the test, asking them to imagine that they were witnesses being questioned by a policeman. They were told that the policeman would provide five alternatives for each question, but since he did not know what happened in the bank, the alternatives would sometimes include the correct answer and other times not. This instruction had two main functions. First, it cancelled out the usual expectation that one of the answers in a multiple-choice test is correct. Second, it served to reduce surprise and interruptions from participants who knew the correct answer but did not see it on the list of five alternatives. The instructions also included a detailed description of the procedure and the different answers that they had to provide (see Appendix A), along with an example question introducing the layout of the experimental section (see Appendix B).

The 28 questions were presented in chronological order. Half of them were presented with the correct alternative and the other half without it. Questions were randomly assigned to each condition and counterbalanced. For each question the participants filled in several sections (see Appendix B). In section A they had to select one alternative (*single answer*) and rate the likelihood that they selected the correct alternative. In section B they had to select three alternatives (*plural answer*) and also rate the likelihood that the correct answer was among their choices. In section C (plurality option) they had to select whether they preferred to answer with one or three alternatives (the *final answer* of the plurality option). Up to this point, the procedure followed the standard plurality option (Luna *et al*., 2011). In section D (report option) the participants indicated whether they preferred to answer with their

selection in section C (*reported answer*) or leave the question unanswered (*withheld answer*). Finally, in section E they were prompted to write the correct answer if they were certain they knew it and it was not among the alternatives. The rationale for this question was that a participant would provide answers at random when she was certain that she knew the correct answer and it was not listed as an alternative. When a participant wrote something in this section, regardless of whether the answer was correct, the data for that question and participant were removed from all of the analyses. This occurred for 10% of all answers, and happened more often when the correct alternative was not presented (16% of the answers to questions without the correct alternative) than when it was (4% of the answers to questions with correct alternative), $t(67) = 7.81$, $p < 0.001$, $d = 1.20$. In support of participants somehow answering at random, we found that 46.3% of the times an answer in section E was accompanied with a decision to leave the question unanswered in section D. Answers were compulsory in all but section E. Finally, the participants were debriefed and dismissed.

RESULTS

We computed and report pairwise comparisons with the Student's *t* test. We also report Cohen's *d* as a measure of effect size and 95% confidence intervals. As confidence is relevant for the single-plural decision, we first report the analyses of confidence and then the proportion of responses. Then we present exploratory analyses to test whether it is possible to classify a question as either presented with or without the correct alternative. Finally, we present the analyses of the memory benefit of the plurality and report options.

Confidence

Main statistics are displayed in Table 1. For single answers, confidence was higher for questions with than without the correct alternative, $t(67) = 8.80$, $p < 0.001$, $d = 0.76$. For the final answer, confidence was also higher with than without the correct alternative, $t(67) = 7.80$, $p < 0.001$, $d = 0.62$. The same pattern was found for single and reported answers, $t(66) = 6.87$, $p < 0.001$, $d = 0.66$.

Proportion of responses

The results of the proportion of responses followed those of confidence. Main statistics are displayed in Table 2. The proportion of plural and reported answers was included for completeness. The proportion of selection of single answers was higher for questions presented with the correct alternative than without, $t(67) = 7.46$, $p < 0.001$, $d = 0.70$. Similarly, the proportion of reported answers was also higher for questions with than without the correct alternative, $t(67) = 4.44$, $p < 0.001$, $d = 0.43$. In addition, the proportion of selection of both single and report together was also higher for questions presented with than without the correct alternative, $t(67) = 9.40$, $p < 0.001$, $d = 0.95$.

Classification of questions with and without the correct alternative

To test whether S100R answers (single answers selected in the plurality option, rated with confidence 100, and reported) have the potential to distinguish between questions with and without the correct alternative, we first computed for each participant the Goodman-Kruskal gamma correlation (gamma) between the number of S100R answers and the type of question, with or without the correct alternative. Gamma was high and greater than zero, $M = 0.55$, $SD = 0.42$, CI [0.45, 0.65], $t(64) = 30.03$, $p < 0.001$, $d = 3.68$, suggesting that S100R answers can distinguish between questions presented with and without the correct alternative.

We then tested whether it was possible to classify *a posteriori* a response as presented either with or without the correct alternative. To this end, we computed the total number of S100R answers for all the sample (361, or 21% of the total number of responses), and the number of S100R answers for questions with correct alternative (259, 72% of all the S100R answers). Note that we did not compute the number of answers per participant and averaged. Here we were interested in specific answers, not in aggregated data. The above proportions mean that a single answer with a confidence rating of 100 and reported has a 72% chance of coming from a question presented with the correct alternative. This proportion was significantly higher than 0.50, $Z = 10.47$, $p < 0.001$.[1]

Memory benefits

In this section, only the questions presented with the correct alternative were analyzed. The memory benefit of the plurality option is the increase from the accuracy for all the single answers ($M = 0.51$, $SD = 0.12$, CI [0.48, 0.54]) to the accuracy for the final answers ($M = 0.69$, $SD = 0.14$, CI [0.66, 0.73]). The increase was indeed significant and with a large effect size, $t(67) = 16.67$, $p < 0.001$, $d = 1.40$. To test the memory benefit of the report option, we compared the accuracy for the final answers and the accuracy for the reported answers ($M = 0.72$, $SD = 0.14$, CI [0.68, 0.75]), $t(66) = 1.74$, $p = 0.087$, $d = 0.18$. The comparison could be considered as marginally significant, but the very low effect size suggests that if there is any effect, it is very small. In sum, the plurality option largely increased accuracy, but the report option did not.

One explanation why the report option may have not increased accuracy is because it did not work as expected, that is, participants failed to discriminate between answers with high and low chances of being correct. To test this idea, we computed the accuracy for

Table 1. *Mean (standard deviation) [95% confidence interval] of confidence for questions presented with and without the correct alternative for Experiments 1 and 2*

	With CA	Without CA	Cohen's *d* *
Experiment 1			
Single	65.83 (12.13) [62.94, 68.71]	55.51 (15.04) [51.94, 59.08]	0.75
Final answer	76.84 (11.36) [74.14, 79.54]	68.69 (14.84) [65.16, 72.21]	0.62
Single and Reported	79.98 (12.72) [76.94, 83.03]	70.40 (16.26) [66.51, 74.30]	0.66
Experiment 2			
Single	45.88 (17.56) [41.06, 50.70]	32.17 (17.59) [27.35, 36.99]	0.77
Final answer	54.98 (21.03) [49.22, 60.74]	40.10 (20.81) [34.40, 45.80]	0.70
Single and Reported	87.30 (16.13) [82.60, 92.00]	78.56 (17.15) [72.09, 85.03]	0.39

Notes: CA: correct alternative. *Cohen's *d* of the difference between with and without the correct alternative.

Table 2. *Mean (standard deviation) [95% confidence interval] of the proportion of selections for questions presented with and without the correct alternative for Experiments 1 and 2*

	With CA	Without CA	Cohen's *d* *
Experiment 1			
PO – proportion of single answers	0.45 (0.18) [0.41, 0.49]	0.32 (0.20) [0.27, 0.37]	0.70
RO – proportion of reported answers	0.68 (0.19) [0.64, 0.73]	0.59 (0.23) [0.54, 0.65]	0.43
Proportion of single and reported	0.41 (0.17) [0.37, 0.45]	0.25 (0.15) [0.22, 0.29]	0.95
Proportion of plural and reported	0.28 (0.19) [0.24, 0.32]	0.34 (0.23) [0.28, 0.40]	0.30
Experiment 2			
PO – proportion of single answers	0.38 (0.27) [0.30, 0.46]	0.24 (0.27) [0.16, 0.32]	0.51
RO – Proportion of reported answers	0.38 (0.22) [0.32, 0.44]	0.25 (0.24) [0.19, 0.31]	0.54
Proportion of single and reported	0.28 (0.19) [0.22, 0.34]	0.13 (0.18) [0.07, 0.19]	0.77
Proportion of plural and reported	0.11 (0.12) [0.07, 0.15]	0.13 (0.17) [0.09, 0.17]	0.15

Notes: PO: plurality option; RO: Report option; CA: correct alternative. *Cohen's *d* of the difference between with and without the correct alternative.

the reported answers (*M* = 0.72, *SD* = 0.14, *CI* [0.68, 0.75]) and for the withheld answers from section D (*M* = 0.47, *SD* = 0.25, *CI* [0.41, 0.53]). Accuracy was higher for reported than for withheld answers, meaning that the report option filtered out answers with low accuracy, $t(61) = 6.76$, $p < 0.001$, $d = 1.30$. However, the report option failed to significantly increase accuracy, most likely because most of the memory benefits that may be obtained with the report option were already obtained with the plurality option.

DISCUSSION

The results of Experiment 1 confirmed that participants' confidence was lower for questions presented without than with the correct alternative, which in turn affected the proportion of selection of single/plural and report/withhold answers. The proportion of selection of both types of answers and the confidence of the single answers offered a potential measure to classify a question as presented with and without the correct alternative. Classification was significantly higher than chance, but our conclusion is that S100R answers may not be good enough in a forensic scenario, because 28% of the S100R answers would have been incorrectly classified as coming from a question with the correct alternative. However, our results suggest that S100R answers have potential for classification.

Another interesting result is that the memory benefits of both plurality and report option overlap. Even though participants were able to use the report option to filter out answers with low chances of being correct, the usual memory benefit was not replicated, probably because most of the memory benefit that can be obtained from the strategic regulation of the answer was already obtained with the plurality option.

EXPERIMENT 2

This experiment was a replication of Experiment 1 with different materials and the addition of a manipulation of the instructions. The objective of the instructions was to increase the number of plural answers. If participants produce fewer S100R answers, the quality of these answers may increase and thus lead to better classification between questions with and without the correct alternative. A secondary objective was to test the generalizability of the results of Experiment 1 using new materials and sample.

METHOD

Participants and design

This experiment was conducted at a Portuguese university, and in Portuguese. Fifty-seven participants completed the experiment. Responses were collected in paper and pencil format and six participants did not provide any answer in Section C (single-plural selection). These participants were removed from the analyses. Thus, data from 51 participants (35 females, mean age 29.22 years old, *SD* = 12.02) were used for the analyses.

The design was the same as in Experiment 1, with the addition of the manipulation of the instructions. Twenty-five participants received instructions similar to those used in Experiment 1, and 26 participants received instructions that highlighted the relevance of providing answers that included the correct alternative. However, the instructions did not affect the proportion of single and plural answers; in fact, they did not affect any of the measures collected and computed. Therefore, we report the results with both groups collapsed.

Materials and procedure

A slideshow with 18 slides showing the discovery of a large knife and a dead body in the bathroom of an apartment was used. Each slide was presented for two seconds separated by a black screen for one second. The slideshow and the questions were developed for a previous study conducted in our laboratory. The questionnaire followed the same sections as in Experiment 1 and included 24 questions, half presented with and half without the correct alternative, counterbalanced.

Participants wrote something in the control question (Section E) in 8% of all the answers. This happened more often for questions without the correct alternative (14% of the answers to questions without the correct alternative) than with (2% of the answers to questions with correct alternative), $t(50) = -5.45$, $p < 0.001$, $d = 0.90$. Also replicating the result of Experiment 1, 45.9% of the answers in section E were made after a decision to leave the question unanswered in section D.

RESULTS

Despite the changes in materials, language, and country of origin of the sample, the results largely replicated those in Experiment 1.

Confidence

Main statistics are presented in Table 1. For single answers, confidence was higher for questions with than without the correct alternative, $t(50) = 7.10$, $p < 0.001$, $d = 0.77$. For final answers, confidence was also higher with than without the correct alternative, $t(50) = 7.61$, $p < 0.001$, $d = 0.70$. For single and reported answers the results were not as clear, but the effect size suggests that there is a small difference, $t(25) = 1.73$, $p = 0.097$, $d = 0.39$. Notice also that for this last analysis many participants were lost because they did not provide any single reported answer, mostly for questions without the correct alternative.

Proportion of responses

Main statistics are presented in Table 2. The proportion of responses of single answers, of reported answers, and of single and reported answers was higher for questions presented with the correct alternative than without, $t(50) = 4.97$, $p < 0.001$, $d = 0.51$, $t(50) = 4.60$, $p < 0.001$, $d = 0.54$, and $t(50) = 5.10$, $p < 0.001$, $d = 0.77$, respectively.

Classification of questions as either with or without the correct alternative As in Experiment 1, gamma was high and higher than zero, $M = 0.71$, $SD = 0.53$, CI [0.53, 0.89], $t(35) = 20.19$, $p < 0.001$, $d = 3.29$. The participants produced 106 S100R answers (9% of the total), and 89 of them (84%) were produced for questions with the correct alternative. This proportion was higher than 0.50, $Z = 12.38$, $p < 0.001$. The classification here was better than in Experiment 1, $Z = 2.56$, $p = 0.010$.

Memory benefits

In this section only the questions presented with the correct alternative were analyzed. The plurality option increased accuracy from all the single answers ($M = 0.45$, $SD = 0.19$, CI [0.40, 0.50]) to the final answers ($M = 0.72$, $SD = 0.16$, CI [0.68, 0.77]), $t(50) = 10.65$, $p < 0.001$, $d = 1.55$. However, the report option did not show any benefit in accuracy. There were no differences between the accuracy of the final answers and the reported answers ($M = 0.76$, $SD = 0.24$, CI [0.69, 0.83]), $t(42) = 1.09$, $p = 0.281$, $d = 0.17$. Again, this was not due to the report

option not working, because accuracy for reported answers was higher than for withheld answers, $t(42) = 2.29$, $p = 0.027$, $d = 0.41$.

DISCUSSION

In general, the results replicated those from Experiment 1, with the main difference that classification between questions with and without the correct alternative was better here. Another difference is that in Experiment 2 confidence was lower for all types of answers and there were fewer single, reported, and S100R answers than in Experiment 1. These results are consistent with the idea that questions were more difficult in Experiment 2 than in Experiment 1. In support of this, accuracy for all single answers was lower in Experiment 2 than Experiment 1, $t(117) = 2.07$, $p = 0.040$, $d = 0.38$. As confidence was also lower in Experiment 2, fewer answers passed the criterion to provide single (vs. plural) and report (vs. withhold) answers. The lower quantity of S100R answers was accompanied by a better quality of those responses, following the quantity-accuracy trade-off that lies at the core of the plurality and report options (Koriat & Goldsmith, 1996; see also the informativeness-accuracy trade-off, Ackerman & Goldsmith, 2008). In sum, the results led to the not very intuitive conclusion that classification between questions presented with and without the correct alternative may be better when questions are more difficult.

Alternatively, the fewer number of single, reported, and S100R answers may be due to participants applying a stricter response criterion in Experiment 2 than in Experiment 1. There may be several reasons why criterion may have changed, ranging from cultural differences to idiosyncratic characteristics of the materials and questions. However, a stricter criterion is usually accompanied by increased accuracy, and that was not the case. Actually, what caused the lower number of S100R answers in Experiment 2 is of secondary importance to the objectives here. The important point is that with fewer S100R answers the classification seems to be better.

GENERAL DISCUSSION

The main objective of this research was to study the pattern of responses when questions do not include the correct alternative. Our main result was that participants respond differently to questions presented with than without the correct alternative. Participants selected more plural (vs. single), more withhold (vs. report) answers, and were less confident in their responses when answering questions without correct alternative. These results suggest that participants can estimate the chances that a given set of answers include the correct alternative, and use that information to guide the control process in charge of deciding the final response.

The present research extends the current understanding of the strategic regulation of accuracy. The decision to select one type of answer or another is driven by the competing goals of informativeness and accuracy, and research has shown that the decision is made based primarily on the likelihood that the answers are correct (confidence criterion) and their informativeness (informativeness criterion; Ackerman & Goldsmith, 2008; Goldsmith, 2016). The model also states that both criteria are affected by different factors, such as incentives, objectives, or social demands. Any factor affecting the control process will likely have an effect on the criteria and, thus, on the outcome of the regulation of accuracy. Our research showed that the control process is sensitive to the presence or absence of the correct response, thus adding it to the list of factors that affect the strategic regulation of accuracy.

Taking advantage of the susceptibility of the control process to the presence or absence of the correct alternative, we used participants' responses to classify *a posteriori* questions as either presented with or without the correct alternative. Results were encouraging, but in general not sufficiently satisfactory. There is still much room for improvement: "only" 72 (Experiment 1) and 84% (Experiment 2) of S100R answers were correctly classified as coming from a question presented with the correct alternative. There are several reasons that may have limited classification. For example, a participant may have no memory of a particular detail, either because it was not encoded or for any other reason. Sometimes, plausible inferences are made when there is no memory, but this is not always possible. In these cases, performance should be similar for a question with or without the correct alternative, because none of the alternatives presented have a match on memory. However, with no memory of the detail it does not seem likely that participants will produce an S100R answer.

Another reason is that sometimes a participant may access an incorrect answer with high confidence, thus producing an S100R answer for a question that may have been presented without a correct answer. For example, after the question "The island of Corsica belongs to what country?" many people may answer Italy, when the correct answer is France (see the consensuality principle, Koriat, 2008). We made sure that our questions were not deceptive, in the sense that there was not a preferred but incorrect alternative; therefore, a systematic bias is unlikely. Another source of incorrect S100R answers is that sometimes an alternative may be activated by processes other than recollection (e.g., inferences made from the situation, or high- typicality information activated from a schema, Luna & Migueles, 2008). In sum, for a variety of reasons sometimes a participant may have accessed an incorrect answer with high confidence, producing an S100R answer. If this happened for a question without the correct alternative, then classification was impaired.

We used a deliberately simple strategy to classify questions with and without the correct alternative from S100R answers. We analyzed individual answers and proportions for the entire dataset. However, other strategies of analyses may have provided interesting information. For example, the computation of the total number of S100R answers correctly classified by each participant (i.e., aggregated analysis) can answer the question of whether some of them can perfectly discriminate between questions presented with and without the correct answer. Indeed, 34% (Experiment 1) and 72% (Experiment 2) of participants had perfect classification. The result is intriguing and suggests that individual differences may play a role, but it is also misleading. Many of the participants with perfect discrimination, particularly in Experiment 2, had only one or two S100R answers. Perfect classification in these cases is not highly informative.

Even though this is an interesting result, this type of aggregated

data is of limited value if we consider the potential application of this research. Telling police that an S100R answer has around a 70% chance of coming from a question with a correct alternative seems more relevant than telling them that participants correctly classify around 70% of the questions with and without the correct alternative (for a similar argument involving correlations and calibrations, see Brewer & Wells, 2006; Luna & Mart´ın-Luengo, 2012a). Thus, although the use of S100R answers as a classification index to distinguish between questions with correct and all incorrect alternatives looks promising, more research is needed for this to become a usable tool for police investigators. For example, researchers may want to add factors associated with memory performance to try to improve classification (e.g., response latency, Sporer, 1992; Weber & Brewer, 2006).

A potential limitation to the application of this research is that we employed a procedure with no widespread use in a forensic scenario. Multiple-choice questions may not be the preferred option during a forensic interview, although Rivard, Pena and Schreiber Compo (2015) found 8% of multiple-choice questions during an interview, suggesting that they are not completely out of place in that context. While our procedure guaranteed the necessary experimental control and comparability with previous studies, future research on this topic should search for ways to improve its ecological validity and the generalization of the findings. For example, witnesses may be asked to generate their own alternatives. More realistic video presentation could also be used with the help of current technologies, for example using 360-degree videos and virtual reality devices, with incidental exposition. This research provides initial insight into the potential interest and relevance of representing the interviewers' lack of knowledge and introduces a way to experimentally study it.

Our second objective was to study the memory benefits when both plurality and report option were available. We found that the memory benefit of the plurality option and the report option overlaps. This is no surprise, as both options share theoretical explanations and are based on the same accuracy-informativeness trade-off (Goldsmith, 2016). However, to our knowledge, this provides the first empirical support for the overlapping memory benefits.

The plurality option increased accuracy, but the report option did not. It would be wrong to interpret these results as suggesting that the plurality option is more powerful, that its memory benefits are greater, or that it is in any way better than the report option. This research did not compare plurality and report options, so no conclusion can be made in that regard. Remember that the plurality option was presented first, so it is reasonable that most of the memory benefit that could be obtained through the strategic regulation of both options was already obtained in the first option, with little room for improvement for the second (the report option). The conclusion from this research is that if improving accuracy is the main objective, there is little benefit in offering both the plurality and the report options sequentially. The plurality option alone can provide most of the memory benefit that can be obtained with the regulation of accuracy. The question remains whether the report option alone may provide similarly high memory benefits.

In summary, this research showed that the control process is

sensitive to the presence or absence of the correct alternative, which adds to the factors affecting the regulation of accuracy (Ackerman & Goldsmith, 2008; Koriat & Goldsmith, 1996; Higham, 2007; Horry *et al*., 2016; Luna *et al*., 2011). The report option seemed to be less sensitive than the plurality option, but that may be the consequence of the specific procedure employed here. Future research should be able to confirm whether there is any advantage of one regulatory strategy over the other.

This article was written while the first author was a visiting researcher at the Centre for Cognition and Decision Making, National Research University-Higher School of Economics (Russian Federation). This study was partially conducted at the Psychology Research Centre (PSI/ 01662), University of Minho, and supported by the Portuguese Foundation for Science and Technology and the Portuguese Ministry of Science, Technology and Higher Education through national funds and co-financed by FEDER through COMPETE2020 under the PT2020 Partnership Agreement (POCI-01-0145-FEDER-007653). This work was also partially supported by the Russian Academic Excellence Project '5-100'.

NOTE

[1] We also tried more sophisticated analyses, including binomial logarithmic regressions. Results did not show any significant improvement over those reported here, which are simpler to interpret.

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.

Brewer, N., Keast, A. & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied*, *8*, 44–56.

Brewer, N. & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*, 11–30.

Goldsmith, M. (2016). Metacognitive quality-control processes in memory retrieval and reporting. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 357–385). Oxford: Oxford University Press.

Herrington, R. (Director) (2002). *The stick-up* [Motion picture]. Universal City, CA: Universal Pictures Video.

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. (2013). Regulating accuracy on university tests with the plurality option. *Learning and Instruction*, *24*, 26–36.

Higham, P. A., Luna, K. & Bloomfield, J. (2011). Trace-strength and source-monitoring accounts of accuracy and metacognitive resolution in the misinformation paradigm. *Applied Cognitive Psychology*, *25*, 324–335.

Hollins, T. J. & Weber, N. (2017). Evidence of a metacognitive benefit to memory? *Memory*, *25*, 317–325.

Hollins, T. S. & Perfect, T. J. (1997). The confidence-accuracy relation in eyewitness event memory: The mixed question type effect. *Legal and Criminological Psychology*, *2*, 205–218.

Horry, R., Brewer, N. & Weber, N. (2016). The grain-size lineup: A test of a novel eyewitness identification procedure. *Law and Human Behavior*, *40*, 147–158.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 945–959.

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

Lindsay, D. S., Read, J. D. & Sharma, K. (1998). Accuracy and confidence in person identification. The relationship is strong when witnessing conditions vary widely. *Psychological Science*, *9*, 215–218. Luna, K., Higham, P. A. & Mart´ın-Luengo, B. (2011). Regulation of memory accuracy with multiple answers: The plurality option. *Journal of Experimental Psychology: Applied*, *17*, 148–158.

Luna, K. & Mart´ın-Luengo, B. (2012a). Confidence-accuracy calibration with general knowledge and eyewitness memory cued recall questions. *Applied Cognitive Psychology*, *26*, 289–295.

Luna, K. & Mart´ın-Luengo, B. (2012b). Improving the accuracy of eyewitnesses in the presence of misinformation with the plurality option. *Applied Cognitive Psychology*, *26*, 687–693.

Luna, K. & Mart´ın-Luengo, B. (2017). The effect of emotional arousal in the regulation of accuracy in eyewitness memory. Manuscript submitted for publication.

Luna, K., Mart´ın-Luengo, B. & Brewer, N. (2015). Are regulatory strategies necessary in the regulation of accuracy? The effect of direct- access answers. *Memory & Cognition*, *43*, 1180–1192.

Luna, K. & Migueles, M. (2008). Typicality and misinformation: Two sources of distortion. *Psicol´ogica*, *29*, 171–188.

Pezdek, K., Lam, S. T. & Sperry, K. (2009). Forced confabulation more strongly influences event memory if suggestions are other-generated than self-generated. *Legal and Criminological Psychology*, *14*, 241– 252.

Pezdek, K., Sperry, K. & Owens, S. (2007). Interviewing witnesses: The effect of forced confabulation on event memory. *Law and Human Behavior*, *31*, 463–478.

Rivard, J. R., Pena, M. M. & Schreiber Compo, N. (2015). “Blind”

interviewing: Is ignorance bliss. *Memory*, *24*, 1256–1266.

Sauer, J. D., Brewer, N. & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General*, *137*, 528–547.

Saulerland, M., Sagana, A. & Sporer, S. L. (2012). Assessing nonchoosers’ eyewitness identification accuracy from photographic showups by using confidence and response times. *Law and Human Behavior*, *36*, 394–403.

Scoboria, A., Mazzoni, G. & Kirsch, I. (2008). “Don’t know” responding to answerable and unanswerable questions during misleading and hypnotic interviews. *Journal of Experimental Psychology: Applied*, *14*, 255–265.

Smith, S. M., Lindsay, R. C. L. & Pryke, S. (2000). Postdictors of eyewitness errors: can false identifications be diagnosed? *Journal of Applied Psychology*, *85*, 542–550.

Smith, S. M., Lindsay, R. C. L., Pryke, S. & Dysart, J. E. (2001). Postdictors of eyewitness errors. Can false identifications be diagnosed in the cross- race situation? *Psychology, Public Policy, and Law*, *7*, 153–169.

Sporer, S. L. (1992). Post-dicting eyewitness accuracy: Confidence, decision-times and person descriptions of choosers and non-choosers. *European Journal of Social Psychology*, *22*, 157–180.

Waterman, A. H., Blades, M. & Spencer, C. (2001). Interviewing children and adults: The effect of question format on the tendency to speculate. *Applied Cognitive Psychology*, *15*, 521–531.

Weber, N. & Brewer, N. (2006). Positive versus negative face recognition decisions: Confidence, accuracy and response latency. *Applied Cognitive Psychology*, *20*, 17–31.

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

Wells, G. L. & Turtle, J. W. (1986). Eyewitness identification: The importance of lineup models. *Psychological Bulletin*, *99*, 320–329.

Received 4 October 2016, accepted 28 March 2017

APPENDIX A

Instructions of Experiments 1 and 2

We would like you to answer some questions about the event you saw at the beginning of the experiment. Imagine that you are a witness of the event and that you are being questioned by a police officer. For each question the police officer will present you five alternatives, but since he does not know what happened, sometimes the alternatives will include the correct answer and sometimes they will not.

Your tasks are:

1. Choose <u>one</u> alternative that you think is most likely to be the correct answer. Indicate this by inserting a number from 1 to 5 in the text box.
2. Indicate the likelihood that this answer is correct. 0% indicates no likelihood at all, and 100% indicates certainty that the option you chose is the correct answer.
3. Choose <u>three</u> alternatives that you think include the correct answer. Indicate this by inserting a number from 1 to 5 in each text box.
4. Indicate the likelihood that the correct answer is *any one of the three alternatives*. 0% indicates no likelihood at all, and 100% indicates certainty that the correct answer is one of the three you chose. Keep in mind that you are more likely to select the correct alternative as you select more alternatives.
5. Suppose you are in Court giving testimony and being asked these particular questions. Select one of the two responses you made (the one-alternative answer or the three-alternative answer) to give the Court.
6. If you were giving testimony in Court, indicate if you would like to give the answer that you selected in the last section of if you would rather leave the question unanswered.
7. If you are sure that the correct answer was not one of the alternatives presented because you know the correct answer, write it down.

APPENDIX B

Example question presented in the instructions. X denoted the selections of a potential participant. The layout during the experimental phase was the same but without the selections.

1. On the day of the murder, what was the weather like?

A. Choose the one alternative that you think is most likely to be correct:

- □ Sunny
- X Light Rain
- □ Overcast, but not raining
- □ Thundershowers
- □ Snowy

Indicate the likelihood that your answer is correct:

o	o	o	o	o	X	o	o	o	o	o
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

B. Choose the three alternatives that you think are most likely to be correct and indicate the likelihood that the correct answer is any one of them:

X Sunny
X Light Rain
X Overcast, but not raining
□ Thundershowers
□ Snowy

Indicate the likelihood that your answer is correct:

o	o	o	o	o	o	o	X	o	o	o
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

C. If you were in Court, your answer would be:
○ A (1 alternative) X B (3 alternatives)

D. If you were in Court, would you like to use your selection in C as an answer or would you rather leave the question unanswered?
X Use C as an answer ○ Leave it unanswered

E. If you are certain that the correct answer was not presented as one of the alternatives, write the correct answer.