Laboratory vs. Naturalistic Prospective Memory Task Predictions: Young Adults Are Overconfident Outside of the Laboratory

Stéphanie Cauvin

Faculté de Psychologie et des Sciences de l'Education, University of Geneva

Christopher J.A. Moulin

Laboratoire de Psychologie et Neurocognition, Université de Grenoble

Céline Souchay

Laboratoire de Psychologie et Neurocognition, Université de Grenoble

Katharina M. Schnitzspahn

School of Psychology, University of Aberdeen

Matthias Kliegel

Faculté de Psychologie et des Sciences de l'Education, University of Geneva

Author Notes:

We thank Mathilde Bastien, Alexandre Caddoux, Asli Erdemli and Vanessa Marti for assistance with data collection. Preparation of this manuscript was funded by a joint grant from the Swiss National Science Foundation (SNSF), the Agence Nationale de Recherche (ANR; France) to MK, CM, KS and CS. CM gratefully acknowledges the support of the Institut Universitaire de France. The authors report no conflicts of interest.

Correspondence concerning this article should be addressed to Stéphanie Cauvin, Boulevard du Pont d'Arve 40, CH-1211 Genève. E-Mail: Stephanie.Cauvin@unige.ch

Abstract

This study investigated whether individuals can predict their future prospective memory (PM) performance in a lab-based task and in a naturalistic task. Metacognitive awareness was assessed by asking participants to give judgments-of-learning (JOLs) on an item-level for the prospective (that something has to be done) and retrospective (what to do) PM component. In addition, to explore whether giving predictions influences PM performance, we compared a control group (without predictions) to a prediction group. Results revealed that giving predictions did not change PM performance. Moreover, participants were underconfident in their PM performance in the lab-based task, while they were overconfident in the naturalistic task. In addition, item-level JOLs indicated that they were inaccurate in predicting what items they will recall or not, but only for the prospective component of the PM task. As for the retrospective component, they were equally accurate in both task settings. This study suggests a dissociation of metacognitive awareness for PM according to both task setting and processing component.

Keywords: prospective memory, metacognition, judgment-of-learning
Words count for the main text: 7’646

Introduction

While sitting at your desk, you think about several items you should buy at the grocery store before heading home. On your way back from work, you pass by the store and this reminds you that you have things to buy. This example illustrates an important memory process that we use everyday: prospective memory (PM; Brandimonte, Einstein & McDaniel, 1996; Kliegel, McDaniel & Einstein, 2008). PM refers to the ability to remember to realize intended actions after a delay either in response to a target event in event-based PM tasks or at a specific time or when a specific amount of time has elapsed in time-based PM tasks (Kvavilashvili & Ellis, 1996). In the above-mentioned example, seeing the store on your way home acts as a target event that triggers retrieval of the intended action (i.e., grocery shopping). A time-based example would be joining a Skype conference call at 3 PM, which necessitates time monitoring to carry out the intention at the appropriate moment.

A particularity of PM tasks is that one has to remember the PM task while being engaged in another (ongoing) activity (Einstein & McDaniel, 1990). To mimic these demands in the laboratory, participants usually perform a computer-based PM task involving an ongoing task (e.g., lexical decision task; LDT) in which the PM task is embedded (e.g., remembering to press a specific key when a predefined target cue occurs among the LDT stimuli). Conceptually, PM has been described as having two process components; the *prospective* component allows one to remember *that* something has to be done (e.g., stopping at the grocery store when passing by), and the *retrospective* component allows one to remember *what* to do (e.g., recalling the list of items to buy; Einstein & McDaniel, 1996).

Besides testing PM in the laboratory, a parallel line of research has explored PM functioning using naturalistic tasks, which were incorporated in the participants' everyday life (for a review see Phillips, Henry, & Martin, 2008). Participants were, for example, asked to remember to send postcards to the experimenter (West, 1988) or to call the laboratory at specific times (Devolder, Brigham, & Pressley, 1990). Importantly, for present purposes, these two research lines have produced conflicting results: A meta-analytic review by Henry, MacLeod, Phillips and Crawford (2004) comparing laboratory versus naturalistic PM performance across young and older adults revealed that young adults performed worse in the field than in the laboratory and even performed worse than older adults in the field. Previous studies suggest that the low performance for naturalistic tasks in the young adults might be explained by a high stress level in their everyday life, an ineffective use of reminders and a low motivation to perform the given tasks (Ihle, Schnitzspahn, Rendell, Luong, & Kliegel, 2012). By contrast, higher performance in these tasks by the older adults may stem from higher motivation and better knowledge of their personal strengths and strategies (i.e., metacognitive awareness, see below; Schnitzspahn, Ihle, Henry, Rendell, & Kliegel, 2011a). In the present paper, we argue that metacognitive factors such as metacognitive monitoring and control may be key when examining performance differences between laboratory and naturalistic PM tasks.

Why add metacognition to the picture? Metacognition refers to the knowledge about our own cognitive abilities. Considering the example of this introduction, may give first insights. Consider that, because you know that it is likely that you will forget the items to buy, you decide to write down the list on a post-it note and to stick it on your computer screen to remind you when you leave work. Monitoring for possible memory failure and thinking of a strategy to prevent it are cognitive processes involved in most PM tasks that have surprisingly been largely overlooked in PM research (but see Gilbert, 2015). Metacognition in memory research is referred to as metamemory (Flavell, 1971) and comprises two sub-processes (Nelson & Narens, 1990): first, we evaluate our own memory performance via monitoring at encoding (e.g., knowing that we will forget our next dentist appointment) and then second, we adjust our behavior in response to the previous evaluation through control (e.g., writing down the date and time of the appointment in our calendar). How is metacognition quantified? In retrospective memory, monitoring at encoding is generally assessed with *judgments-of-learning* (JOLs; Nelson & Dunlosky, 1991). Here, they are usually evaluated by asking participants to learn related (e.g., table – chair) or unrelated pairs of words (e.g., table – bus) and then participants either give a global JOL or item-by-item JOLs. In global JOLs, participants predict how many items they will recall or estimate the percentage of cues they expect to find across the entire task. In item-by-item JOLs, participants estimate in percentage, for each pairs of words separately, the probability of recalling them. Moreover, JOLs can be assessed directly after the encoding of each item (i.e., immediate JOLs) or after the encoding of all the items (i.e., delayed JOLs). Judgments made after a delay are more likely to reflect retention in long term memory and have thus been shown to be more reliable (Rhodes, 2016).

In the retrospective literature, metacognitive performance varies according to the task, the methodology used and the memory domain (for a review see Castel, Middlebrooks, & McGillivray, 2016). As there is not a clear pattern of age-related preservation or decline, metacognitive awareness could differ between the retrospective and prospective domain, as well as between laboratory and naturalistic tasks. What do we know about metacognition for PM? Initial evidence has been reported that, regarding *laboratory* tasks, people showed some insights in their performance but tended to *under*estimate themselves (Knight, Harnett & Titov, 2005; Meeks, Hicks & Marsch, 2007; Schnitzspahn, Zeintl, Jäger & Kliegel, 2011b). In Meeks et al.' (2007) study, participants provided a global JOL on their performance on a laboratory PM task. Results revealed participants to be generally underconfident in their PM performance. However, the predictions correlated positively with PM performance, but only in one out of two cue conditions. Thus, individuals seem are better with predicting performance for certain intentions than others (see Schnitzspahn et al., 2011b, for similar findings).

With respect to *naturalistic* tasks, there is only one early study that has explored this question; interestingly revealing the opposite pattern: here, young adults *over*estimated their PM performance. In a sample of young and older adults, Devolder et al. (1990) asked their participants to give global predictions prior to multiple retrospective memory tasks in the laboratory and to one PM task outside of the laboratory. The latter was an appointment keeping task in which the participants had to call the experimenter at specified times over a 4-week period. Young participants not only overestimated their future PM performance, they also had lower performance and were less accurate in their predictions than the older adults.

Taken together, there is initial evidence that (a) metacognitive awareness can be assessed for PM similarly to retrospective memory and – more importantly – that (b) performance predictions may differ between the laboratory and naturalistic tasks. This task setting difference in metacognitive evaluation of PM could in turn critically influence how participants encode and plan their future intentions across both settings and finally how they perform. Thus, the first aim of our study was to compare individuals' PM performance in two traditional PM task types, that have been widely used in the laboratory and in naturalistic settings and to examine whether assessing performance predictions influenced actual performance in both settings.

As an additional aspect, all participants performed the PM tasks in the two settings, but only half of them gave performance predictions. This design was motivated by the rationale that making predictions could increase the perceived importance of the task (Hering, Phillips, & Kliegel, 2013) or help to anticipate similarly to an implementation intentions strategy (Schnitzspahn & Kliegel, 2009). This could result in better PM performance and/or affect costs in the ongoing task, if participants monitor more for the PM cues. Indeed, one has to consider a potential complication when assessing item-level JOLs for laboratory and naturalistic PM tasks that may be particularly relevant for PM research. In detail, in PM tasks, performance predictions could alter PM performance itself as they may enhance PM performance by repeatedly highlighting the intention after initial encoding for future retrieval (see, e.g., Meier, Wartburg, Matter, Rothen, & Reber, 2011 who showed that performance predictions may make the intention more accessible or Rummel, Kuhlmann & Touron, 2013, who found that PM performance increased in a prediction group compared to a control group without performance predictions).

The second aim was to investigate metacognitive processes in PM in a fine-grained manner by asking participants to predict their performance with predictions on an item level separately for a traditional laboratory and naturalistic tasks. While global JOLs allow to estimate a general score of over- or underconfidence, JOLs that are provided at an item-level allow distinction and interpretation of two aspects in performance predictions, (a) whether individuals are under- or overconfident about their performance by comparing the average JOLs with the percentage of items actually recalled (i.e., calibration of predictions to the performance), and (b) whether individuals can correctly discriminate between recalled and non-recalled items (i.e. resolution of judgments) calculated by the Goodman-Kruskal correlation (referred hereafter as gamma correlation; Nelson, 1984). To extend the level of analysis in that matter as exploratory research questions, we examined predictions (a) on the item level and (b) we disentangled performance and JOLs for the two main process subcomponents in a PM task; i.e., the prospective and retrospective component. So far, only one study has investigated whether item-by-item JOLs in a PM task show the same effects that are found in retrospective memory (Schnitzspahn et al., 2011b).

The authors asked 60 undergraduate students to read a story on a computer as an ongoing task. Prior to reading the story, the participants had to learn ten PM cues which were specific words that were paired with an action (e.g., cafeteria – greet the cook). The cues were evenly distributed in the text and participants were instructed to click on the PM cue (prospective component) and to type the associated action (retrospective component). They were asked to predict their future performance by giving item-by-item JOLs for each word-action pair; separately for the two components, first as the probability of becoming aware that something has to be done upon encountering the PM cue, and second as the probability of retrospectively recalling the content of the associated action. The authors found a positive relationship between predictions and performance for both PM components, with higher correlations for the retrospective component than the prospective one. This result suggests that on an item-level young adults have some awareness in their performance for lab-based future intentions, but that they tended to underestimate their performance. More specifically, participants were underconfident for the prospective component, but overconfident for the retrospective component.

To follow-up on this study, our goal was to further zoom-in on the specific processes underlying the relation between metacognition and PM performance by disentangling predictions for the two main sub-processes of PM: the prospective and retrospective components. This may allow to decompose the source of possible metacognitive over / underestimations and to further specify whether over- versus underestimation may stem from rather attentional or memory-related process characteristics of PM tasks. Specifically, in an event-based laboratory task requiring to remember to press a specific key whenever encountering a specific PM cue performance can fail in two aspects: either one can miss the PM cue (prospective component error) or one can forget what action to perform in response to the prospective cue (retrospective component error). The first aspect has been related to rather attentional and executive functioning (Kliegel, Mackinlay, & Jäger, 2008), while the second aspect has been associated to episodic memory (Einstein & McDaniel, 1990; Smith & Bayen, 2006). To be able to disentangle both components in the present study, performance predictions were made for both components separately, in the laboratory and also for the naturalistic task.

To sum up, we investigated PM performance and performance predictions with several research questions. First, regarding PM performance, we examined possible better PM performance for the prediction group compared to the control group. In addition, by comparing PM performance in a traditional laboratory compared to a traditional field task, we expected to replicate better performance in the laboratory than in the field. Second regarding predictions, we examined calibration of the predictions to the performance on the prospective component and tested whether we could corroborate the general underconfidence for a PM task in the laboratory and overconfidence for a PM task in the field. As additional exploratory research questions, we examined resolution of the judgments in both settings for the first time. Moreover, we examined a possible differentiation in both performance and predictions for the prospective and retrospective components of the tasks expecting better performance, as well as better calibration and higher accuracy for the retrospective component compared to the prospective one.

Method

Participants and Design

In total, 87 young adults (*M* = 22.04 years, *SD* = 2.16; age range = 19-30 years; 11 male) took part in the study.[1] Participants were students at the University of Geneva and participated in exchange for course credits. All participants were native French speakers. The Ethics Committee of the University of Geneva approved the study and all participants gave informed consent prior to taking part in the experiment.

The study followed a 2 x 2 x 2 mixed factorial design to investigate the effects of group, setting and component on PM performance and predictions. The between-subjects variable was *group* (control, prediction) and the within-subjects variables were *setting* (lab-based, naturalistic) and *components* (prospective, retrospective).

Materials

Laboratory-based task

Ongoing task. The lab-based PM task followed the traditional paradigm from Einstein and McDaniel (1990) in which the PM task is embedded in an ongoing task. The latter was a lexical-decision task (LDT). The verbal material came from a pool of 866 French words rated by valence and subjective frequency (Bonin et al., 2003). From this set, we selected 210 words of neutral valence, medium frequency and similar length between one and two syllables. Non-words were created by switching the consonants of each word stimulus, resulting in an equal number of valid French words (e.g., Tonneau) and pronounceable non-words (e.g., Noteau). Each LDT trial (font in 36 point, white color) was presented in the center of a black screen, followed by a fixation cross of variable duration (250-750 ms). The stimuli (i.e., either a word or non-word) stayed on the screen until the participant responded. Participants were instructed to press the “yes” key with the right index finger (j-key) if the stimulus presented was a valid French word and the “no” key with the left index finger (f-key) if the stimulus was not a valid French word. They performed a training session with 10 trials before the test phase. After each trial, a feedback was given as “correct” or “incorrect” and at the end of the training, participants were informed of their score in percentage. Under 80% correct, participants had to redo the training.

PM task. For the PM task, participants were further instructed that we were interested in their ability to perform a second task alongside the LDT. They were asked to remember to press the “white key” (i.e., a key marked in white on the keyboard) instead of the “yes/no” key when they identified one of the pre-specified PM cues. The PM task consisted of 15 cue-word pairs that the participant had to memorize prior to the test phase (i.e., the LDT). For encoding, each PM cue was presented as the first word of a pair and the participants were instructed to learn the cue words and the related words as word pairs (e.g., Collier - Ruche). The cue-word pairs consisted of two non-related words each controlled for having 2 syllables and neutral valence taken from the words set mentioned above (Bonin et al., 2003). During the encoding, the stimuli were presented twice and for 4 seconds each. During the LDT, participants were instructed to press the white key instead of their ongoing task response when they detected one of the PM cues (i.e., the first word of the learned cue-word pairs); when they did so, a text box appeared, and they were asked to type in the related word (i.e., the second word of the word pair; note that they could press the white key on any word stimulus of the task). This approach enabled us to disentangle the two different PM components within PM performance: (a) performance on the prospective component, measured by the accuracy of remembering to press the PM key when encountering the PM cues, and (b) performance on the retrospective component, examined by the accuracy of retrieving the related word to the cue (see Schnitzspahn et al., 2011b; Cauvin et al., 2017, for a similar procedure). Additionally, the participants were instructed to press the PM key whenever they detected a PM cue, even if they already gave the ongoing task response or even if they forgot the word related to the cue. The participants were specifically told that both tasks were important and that they should answer as fast and as accurately as possible. To analyze possible additional costs of the PM task, two versions of the task - an ongoing task only (70 stimuli) and the ongoing task with a PM task embedded (150 stimuli) - were given to the participants in a counterbalanced manner. Participants were explicitly informed before the beginning of the block that they would perform the ongoing task only.

JOLs. For the laboratory task (and the naturalistic task, see below), after memorization of all the cue-word pairs participants were asked to give two types of JOLs for each item (see Figure 1). Participants estimated the probability of successful future PM performance on a scale from 0% to 100% (0%, 20%, 40%, 60%, 80%, or 100%; e.g. Nelson & Dunlosky, 1991). First, a pJOL for the *prospective* component as the probability of becoming aware that something had to be done upon encountering a PM cue and second, an rJOL for the *retrospective* component as the probability of retrospectively recalling the associated word of the specific PM cue. Importantly, in each case the question did not include the specific PM cue, but rather used a generic term. For example, in the lab-based task for the "Train – Papier" cue-word pair, the pJOL was "How probable is it that you will press the white key upon encountering a *vehicle*?", and the rJOL was "How probable is it that you will remember the word associated with a vehicle?". Each cue belonged to a different superordinate category and by using this method we avoided explicitly repeating the specific PM cues after the encoding phase.

Naturalistic task

PM task. The naturalistic task was a traditional paradigm that has been used repeatedly to assess PM in everyday life (especially in the context of the age PM paradox; see e.g., Aberle, Rendell, Rose, McDaniel, & Kliegel, 2010; Schnitzspahn et al. 2011a). Participants were asked to remember to send text messages to the experimenter at specific target times. The times were set in the laboratory together with the participants, one in the morning and one in the afternoon over four days, starting the afternoon after the initial laboratory session (for details on the procedure, see below). The time frame intervals were: in the morning between 9 a.m. and 1 p.m. and in the afternoon between 2 p.m. and 6 p.m. Participants could choose any time that suit them best, the second constraint was that each hour and minutes should be different. For each PM target time, a pre-defined word was associated as the content of the text message (the same words for all participants). Participants had 5 minutes to memorize a total of 7 time-word pairs and were informed only after encoding that they could keep the schedule with the times, but not the words, as a reminder. Messages sent six minutes before or after the prescribed time were classified as correct PM answers (for a similar procedure, see Schnitzspahn et al. (2011a).

JOLs. For the naturalistic task, the pJOLs were adapted for the time-based task so that the question referred to the general time in which the task had to be performed, for example "How probable is it that you will send a text message on Monday afternoon?"; while the rJOL referred to the word associated with the specified time, for example "How probable is it that you will send the correct word on Monday afternoon?".

Procedure

The experiment was divided into two sessions a week apart. To control for any task order effect, half of the participants started with the lab-based task, while the other half started with the naturalistic task. For the same reason, to control for the PM load effect on the computer task, half of the participants started with the ongoing task only, while the other half started with the PM task. The first session started with the consent forms and a sociodemographic questionnaire. Participants who started with the naturalistic task first, came in the laboratory to give their consent and to receive the instructions for the task. They came back one week later for the follow-up on the naturalistic task and the lab-based task. The other participants started with the computer task, received the instructions for the naturalistic task and came back for a follow-up on the experiment one week later. An individual session with the computer task lasted approximately 1 hour, while an instruction only session or a follow-up session only lasted 30 minutes; the two sessions lasting in total one hour and a half.

The procedure described in more detail below, was the one for a participant who started with the lab-based task in the prediction condition; the only difference with the control condition was the extra JOL assessment. Participants were randomly assigned to the prediction condition, so that half of them answered JOLs while the other half did not. In the first session, after informed consent was obtained, participants first completed a socio-demographic questionnaire and then were instructed with the computerized LDT. After succeeding the training trials half of them started either with the ongoing task only, or were given further instructions about the PM task. To ensure that the participants understood the instructions correctly, they were asked to orally explain the task back to the experimenter. All participants were able to correctly explain the task right away and therefore could proceed to the encoding phase. Following the encoding of all the PM cues, participants gave their JOLs for each cue and the related word. Then, there was a 3-minute delay, in which participants performed the Digit-Symbol test before the PM test phase began. At the end of the computer task, as a recognition test, participants were asked to decide among 60 stimuli if yes or no the word presented had been part of the words they had to learn for the test phase (15 cues and 15 related words). Then, they received the instructions for the naturalistic task and scheduled the times to send the text messages as well as the next session with the experimenter. After the naturalistic PM encoding phase, they again gave the JOLs, and only after this were they informed that they could keep the calendar as a reminder. Participants finished the session by completing the vocabulary test, as well as a questionnaire on computer and phone usage. One week later, participants came back for a follow-up on the naturalistic task, consisting in general questions about the task, such as if they used the sheet as a reminder and what they were doing when they had to send the text messages. The follow-up also comprised a recognition test for the 7 related words, they had to decide if yes or no the word presented was part of the words they had to send among 30 stimuli. Participants were then debriefed and thanked for their participation.

Results

Data analyses

In the following result section, two main types of analyses will be done. One type refers to confirmatory analyses, which will test the explicit hypotheses that was derived from the literature (see below). The second type refers to exploratory analyses, which will explore a number of new research questions where we could not derive clear hypotheses from the existing literature. Following our hypotheses and as confirmatory analyses, we tested the effects of JOLs on performance. First, on ongoing task performance with 2 x 2 (group x PM load) repeated measures ANOVA and second, on PM performance with separate independent t-tests. Then, in the prediction group only, we analyzed mean PM performance, mean JOLs and gamma correlations in 2 x 2 (setting x component) repeated measures ANOVAs. As we tested performance and calibration for the prospective and retrospective in the same analyses (as a within-subject factor), it is to be noted that analyses regarding the prospective component were confirmatory, while analyses regarding the retrospective component were exploratory. All of the analyses on JOLs resolution were exploratory.

Testing the effects of JOLs on performance

Ongoing task performance

To test whether performance predictions and an added PM task influenced ongoing task performance, we analyzed ongoing task accuracy and reaction time in 2 *group* (control vs. prediction) x 2 *PM load* (without PM vs. with PM) ANOVAs.

Participants who gave performance predictions had the same ongoing task accuracy (*M* = .91, *SE* = .006) as the participants who did not give predictions (*M* = .91, *SE* = .006), *F*(1,86) < 1. As expected, performance was worse with the PM load (*M* = .87, *SE* = .005) than without (*M* = .95, *SE* = .004), *F*(1, 86) = 250.851 , *p* < .001, η_p^2 = .742. The interaction between group and PM load was not significant, *F*(1,86) < 1.

As for the reaction times, participants who gave performance predictions were faster (*M* = 718ms, *SE* = 18ms) than the ones who did not (*M* = 796ms, *SE* = 19ms), *F*(1, 86) = 8.576 , *p* < .005, η_p^2 = .091. As expected, participants were slower with the added PM load (*M* = 63s1, *SE* = 10ms) than without (*M* = 883ms, *SE* = 18ms), *F*(1, 86) = 398.779, *p* < .001, η_p^2 = .823. The interaction was again not significant, *F*(1,86) = 2.412, *p* = .124.

Prospective memory performance

In the laboratory, the retrospective component (recalling the associated word) is tied to the prospective response in the sense that when participants missed a PM cue, they could not recall the associated word. To account for this, we computed retrospective accuracy based on the PM hits for each participant. This was not necessary for the naturalistic task, since participants could miss the time for the prospective response, but still send the associated word later. On average, this happened 3.5 times (*SD* = 1.9). We also removed the cues that were missed in the recognition test at the end of the experiment for each participant, considered as retrospective failure (mean performance in the lab-based task (*M* = 0.87, *SE* = 0.06)); and naturalistic task (*M* = 0.98, *SE* = 0.02). To allow comparisons between prospective and retrospective performance in both settings and later comparisons with the mean JOLs, we transformed the mean scores of performance as percentages.

To test whether giving predictions (answering JOLs) influenced PM performance, we conducted separate *t* tests on PM scores for the prospective and retrospective components in both laboratory and naturalistic settings (see Table 1 for inferential statistics and descriptives). The differences between the control and the prediction groups did not reach statistical significance (all *ps* > .488). Consequently, the following analyses are conducted in the prediction group only to allow direct comparison between performance and predictions. All data are percentage indicated to facilitate subsequent comparison.

Laboratory versus naturalistic PM performance in the prediction condition

To examine PM performance of the prediction group in the two settings, we analyzed PM scores in a 2 *setting* (laboratory vs. naturalistic) x 2 *component* (prospective vs. retrospective) repeated-measures ANOVA.

Participants reached similar PM performance levels in the laboratory (M = 64.13, SE = 3.63) and in the field (M = 57.92, SE = 2.77), $F(1, 45) = 2.031$, $p = .161$, $\eta_p^2 = .043$. Performance on the retrospective component (M = 72.12, SE = 2.93) was better than on the prospective one (M = 49.93, SE = 2.45), $F(1, 45) = 75.393$, $p < .001$, $\eta_p^2 = .626$. The interaction between setting and component was significant, $F(1, 45) = 33.317$, $p < .001$, $\eta_p^2 = .425$. Follow-up tests revealed that participants had better performance on the prospective component in the laboratory compared to the field, $F(1, 45) = 30.020$, $p < .001$, $\eta_p^2 = .400$. But they had better performance on the retrospective component in the field than in the laboratory, $F(1, 45) = 7.554$, $p = .009$, $\eta_p^2 = .144$. Moreover, performance on the prospective component and on the retrospective one did not differ in the laboratory, F (1, 45) = .116, $p = .735$, $\eta_p^2 = .003$, while performance on the retrospective component was better than the prospective one in the field, $F(1, 45) = 123.141$, $p < .001$, $\eta_p^2 = .732$. Note that a link between the prospective and retrospective component in the laboratory may be due to our paradigm, which was supported by a marginally significant correlation ($r = .282$, $p = .058$, N = 46).

However, in the field, where both components could be assessed independently, there also was a marginally positive relation between the two components (r = .277, p = .063, N = 46).

Exploring the predictions

JOL calibration

The first analytical step in exploring performance predictions consists in estimating, in the prediction group, participants' calibration by looking into JOLs magnitude (mean predictions) in order to compare it with mean performance. When participants missed a PM cue they could not recall the associated word, thus we left out the predictions made on the retrospective component (rJOLs) linked to the PM misses. However we considered all the predictions for the prospective components (pJOLs) in both settings and the rJOLs in the naturalistic tasks. To test if predictions follow the same pattern as PM performance, we analyzed participants' predictions in a 2 *setting* (laboratory vs. naturalistic) x 2 *component* (prospective vs. retrospective) ANOVA.

Participants predicted lower performance in the lab-task (M = 48.22, SE = 2.86) compared to the naturalistic task (M = 61.27, SE = 2.46), F(1, 45) = 15.357, p < .001, η_p^2 = .254. Hence showing underconfidence for the lab-task and overconfidence for the naturalistic task, considering that mean performance was comparable across settings. Participants gave similar predictions for the prospective (M = 54.15, SE = 1.70) and the retrospective components (M = 55.34, SE = 2.89), F(1, 45) = .275, p = .603, η_p^2 = .006. However the interaction between setting and component was significant, F(1, 45) = 33.570, p < .001, η_p^2 = .427. Follow-up tests revealed that in the lab-task, participants gave higher predictions for the prospective component compared to the retrospective one, F(1, 45) = 12.457, p = .001, η_p^2 = .217, while in the naturalistic task participants gave higher predictions for the retrospective component compared to the prospective one, F(1, 45) = 11.454, p = .001, η_p^2 = .203. They also predicted better performance for the retrospective component in the field compared to the laboratory, $F(1, 45) = 45.195$, $p < .001$, $\eta_p^2 = .501$; following the same pattern as the performance. However, they predicted the same prospective performance in the laboratory and in the field, $F(1, 45) = .748$, $p = .392$, $\eta_p^2 = .016$.

For a global indication of over- or underconfidence, we compared the mean JOLs with the actual mean performance (all *ps* < .001), by looking at the difference between the two (see Figure 2 for difference scores, all scores reliably greater than zero, *ps* < .001). The pJOLs and rJOLs in both settings were positively related, with a high correlation in the laboratory ($r = .719$, $p < .001$, $N = 46$) and a medium one in the field ($r =.428$, $p < .001$, $N = 46$).

JOL resolution

The second analytical step in exploring performance predictions consists in estimating participants' resolution by looking into Gamma correlations between each participant's JOL (Nelson, 1984). Gamma is a continuous variable that ranges from -1 to +1. A large positive value means a high degree of accuracy; a value of zero means chance-level accuracy, and a negative value means less than chance-level accuracy. To examine if relative accuracy differed between task settings, we analyzed participants' Gamma correlation in a 2 *setting* (laboratory vs. naturalistic) x 2 *component* (prospective vs. retrospective) ANOVA.

Participants showed the same resolution in predicting their performance in the laboratory ($M = .230$, $SE = .053$) and outside of the laboratory ($M = .209$, $SE = .042$), $F(1, 45) = .101$, $p = .753$, $\eta_p^2 = .002$. Gamma correlations did not differ between the prospective ($M = .154$, $SE = .056$) and the retrospective component ($M = .285$, $SE = .042$) of the JOLs, $F(1, 45) = 3.252$, $p = .078$, $\eta_p^2 = .067$. However the interaction between setting and component was again significant, $F(1, 45) = 11.677$, $p < .01$, $\eta_p^2 = .206$. Follow-up tests showed that in the lab-task Gamma correlations did not differ between the prospective ($M = .288$, $SE = .067$) and the retrospective component (*M* = .173, *SE* = .074), *F*(1, 45) = 1.558, *p* = .218, η_p^2 = .033, while in the naturalistic task Gamma correlations were higher for the retrospective (*M* = .398, *SE* = .055) than the prospective component (*M* = .020, *SE* = .082), *F*(1, 45) = 11.536, *p* = .001, η_p^2 = .204. Gamma correlation for the prospective component was better in the laboratory than in the field, *F*(1, 45) = 7.385, *p* = .009, η_p^2 = .141. By contrast, gamma correlation for the retrospective component was better in the field than in the laboratory, *F*(1, 45) = 5.142, *p* = .028, η_p^2 = .103.

Overall mean Gamma correlations were reliably greater than zero for both components in the lab task (Gamma pJOL: *t*(45) = 4.333, *p* < .001; Gamma rJOL: t(45) = 2.332, *p* =.024) and the retrospective component in the naturalistic task, *t*(45) = 7.27, *p* < .001); indicating above-chance resolution of JOLs. However, mean Gamma correlation for the prospective component in the naturalistic task did not differ from zero, *t*(45) = .247, *p* = .806, indicating a chance-level accuracy.

Discussion

The first aim of this study was to examine whether people can accurately predict their future PM performance in laboratory and/or naturalistic contexts and whether making predictions would influence performance by comparing a control group and a prediction group. As a second aim we examined these predictions on a fine grained level using item level predictions and explored whether the predictions may differ for PM task components, disentangling a prospective and a retrospective components of each PM task.

The influence of performance predictions

Previous studies have found that global JOLs may influence PM performance positively (Meier et al. 2011; Rummel et al., 2013). Our results did not replicate this effect and showed that PM performance in the prediction group was comparable to the control group. This is likely related to the way our JOLs were introduced. Although giving predictions for each item could have enhanced activation of the PM task in mind, the use of superordinate categories instead of repeating exactly the specific cue words may have prevented an additional encoding compared to the control group. Moreover, the delay interval between the JOLs and the PM task with a cognitively demanding task prevented the PM task from being too present in mind (Schnitzspahn et al., 2011b). However, it has to be noted that although giving predictions did not change overt PM performance, it seemed to have somewhat influenced ongoing task performance for the PM cues: participants in the prediction group were faster in the LDT than the control group, while showing the same accuracy. This may be a sign of more automatic processing. Thus, future research is needed to investigate the possible influence of JOLs on monitoring of cues. However, compared to former studies with global JOLs, the present paradigm with item-by-item JOLs using generic categories did not directly influence the PM results.

Predictions of future PM performance: Exploring underlying process components

The pattern of performance in the two different contexts was reflected in the predictions. In fact, the participants gave higher pJOLs than rJOLs in the laboratory and higher rJOLs than pJOLs in the field, thus showing some metacognitive awareness of their future PM performance. With regard to the prospective component of PM performance we replicated what is usually found in young adults: PM performance was better in the lab-based task compared to the naturalistic task (Henry et al., 2004). Interestingly, participants performed better on the prospective component than the retrospective one in the laboratory, by contrast they performed better on the retrospective component than the prospective one in the field. More specifically, in the lab-based task participants were better at recognizing PM cues than at retrieving the associated word. In the naturalistic task, participants performed relatively poorly in sending the text messages on time, but were better at retrieving the associated word; even remembering the content after a three-day period. This is a typical PM failure: we often realize too late that we had something to do, but we do not forget what it was. However, this result needs to be interpreted with caution as not only the task setting was different but also the task type (event-based vs. time-based). Future research needs to use the same type of tasks to observe if there is a possible double dissociation with respect to setting and PM components.

Regarding the predictions, participants seemed not to be aware that the two tasks differ in nature, but they seem to differenciate the two components of a PM task. However, the magnitude of the JOLs was not perfectly in line with the actual performance. In the laboratory, participants predicted poorer performance on the prospective component, while in the field they gave the same magnitude of predictions as in the laboratory, but performed less well on the PM task. In previous studies, individuals tended to underestimate their performance in an unfamiliar lab-task (Meeks et al., 2007, Schnitzspahn et al., 2011b), which was also found here. Interestingly, in contrast to this, the same participants were overconfident in the naturalistic setting which dovetails with Devolder et al.'s (1990) findings. In terms of performance, this overconfidence might have led to a failure in generating appropriate strategies to implement the delayed intention; suggesting a failure in the control process of metacognition. Indeed, the obtained lack of a benefit of making predictions on performance seems to be in line with this interpretation. Moreover, in the laboratory predictions for the retrospective component were similar to the prospective one suggesting that participants may have used the same criterion for retrospective and prospective processes, not realizing that remembering the initiation of intentions may be different from traditional declarative memory retrieval. In contrast, in the field, there was a considerable difference between the predictions on the prospective and retrospective components. This suggests that on this task, participants may have been aware of the difference between remembering to perform the task in time and retrieving the word from long term memory. Again, this explanation needs to be further explored to disentangle whether the difference in predictions are due to the setting or the nature of the PM task.

In the light of our exploratory research questions, important new insights are further provided by considering the accuracy of the item-by-item predictions. This measure shows if participants correctly discriminated between items they will remember or not. Here, we observed that participants were as accurate for the prospective as for the retrospective component in the laboratory, but they were only accurate on the retrospective component in the field and were at chance-level for the prospective component. They were not only overconfident in their predictions, they were also inaccurate in predicting whether they will send the text message or not in that their predictions did not discriminate between actions that were ultimately carried out on time or forgotten. These findings dovetail with Schnitzspahn et al. (2011b) who had found that the accuracy of the retrospective component was more similar to what is usually obtained in laboratory studies of retrospective metamemory using standard word pair tasks. The present study further confirms that it seems more difficult to predict whether we will recognize the target or send a text message than to predict retrospective retrieval. To forget a future intention could have severe consequences in everyday life, but it seems that people have difficulties in predicting if their PM will fail or not, which could, in turn, prevent them from generating useful strategies.

One possibility is that monitoring of what we have learned, for example a word list, a set of facts, or a phone number, is a form of metacognition that is better trained and more habitually used than the monitoring of future intentions. For instance, repeated study-test procedures in education mean that we become accustomed to what are reasonable levels of performance for verbal materials. Moreover, the JOL procedure for learned materials (in our task operationalised as an rJOL) actually includes the possibility of self-test in a task-relevant manner, which is not possible in the pJOL condition. Numerous studies have identified a delayed JOL effect in retrospective memory (e.g. Nelson & Dunlosky, 1991), which suggests that people use a retrieval attempt to inform their JOL prediction. In our task, participants had five minutes to encode the items and then made their predictions. It is possible, that - to make these rJOLs - participants assessed whether they can recall the target word, and this acted as a 'practice': here, the JOL procedure included a task relevant retrieval of the to-be-remembered item. In contrast, in making a pJOL, any retrieval attempt is not going to be diagnostic for future PM performance: one can imagine one's performance based on a number of cues and factors, but one does not have the same possibility to self-test. The fact that we found rJOLs and pJOLs to be correlated seems to suggest that participants misapplied their traditional memory heuristics and that cues they used for rJOLs were also used in making their pJOLs.

Another *post hoc* explanation could be that individuals do not consider or appreciate all the factors involved in PM tasks. This is particularly the case in the naturalistic task, where the task is unconstrained and distractions and impediments to performance are potentially infinite. First, for instance, the duration and thus the delay of the naturalistic task was longer than in the laboratory, and lasted over three days. Second, people may have given optimistic predictions based on their optimal performance, without considering that in everyday life unexpected events can occur that will prevent them to execute the intention as planned.

Another possible explanation for discrepant metacognitive predictions compared with actual performance is the relative difficulty of the two tasks. To stay as close as possible to the traditionally used PM paradigms, we used two well established and often applied PM tasks but are aware that one could argue that they are not comparable in nature (e.g., event versus time), even if they are usually directly compared in the context of research on the age PM paradox (Henry et al., 2004; Schnitzspahn et al., 2011b). Thus, it would be compelling in future studies to use laboratory and naturalistic tasks that are more similar and to compare them regarding performance as well as predictions (e.g., two time-based tasks). A challenge in creating a truly naturalistic task is to maintain experimental control. However, it would be interesting to test if the effects revealed in the present study hold in real-life tasks. This might be possible with a diary keeping approach (see Schnitzspahn et al., 2016).

It should be noted that the sample is predominantly female and university students in Psychology. Thus, the results regarding performance or metacognition may not be generalizable to other institutions or demographics. Future studies should include larger ranges in the level of education of participants, as well as extend the sample to older participants to look into age-related changes in performance and metacognitive awareness.

Finally, in the metamemory literature, there has been very little examination of metacognitive processes in task delays applied in the present study rendering our study unique in that sense (but see Weaver et al. 2008, for differences in predictions on laboratory and real world materials). Thus, it appears important to underline that our results suggest that moderately accurate JOLs are possible for the retrospective component of a task even over a four-day period. In other words, predictions made after studying items for five minutes can be accurate days later (if anything, the rJOLs were even more accurate for the naturalistic task than for the laboratory task).

Our study offers important first insights in key questions such as how accurately individuals predict their future intentions in different contexts and if these predictions influence performance, it also leaves several open issues. We acknowledge that our explanations are partly preliminary and have to remain somewhat speculative in some questions. Future studies looking into PM and metacognition should also consider the control process of metamemory. Indeed, over- or underconfidence is likely to affect what we will do to aid future memory retrieval. In terms of future directions, we suggest that studies on metacognition in PM could bring answers to unresolved questions in the field of PM.

Overestimation – in our case by young adults - in naturalistic tasks could explain why they perform less well than older adults in this context. To test this, future studies should address predictions in young and older adults in both contexts. This could give some insights in the “age-PM paradox” (Henry et al., 2004; Rendell & Thomson, 1999; Schnitzspahn et al., 2011a).

Footnote

[1] Sample size had initially been calculated in the context of a study design comparing young and older adults. Here, an a priori power analysis indicated that a total sample size of N = 176 (88 young and 88 older adults) is large enough to detect a medium effect of ɳ2 = .06 (f = .25) with an alpha probability of .05 and a power of .90 (all power analyses were conducted using G*Power 3.10). Due to administrative reasons, the older cohort could not be tested with the present protocol and the present paper therefore focuses on the younger cohort only. Note, that we tested 90 younger participants altogether but had to exclude three participants because their native language was not French.

References

Aberle, I., Rendell, P. G., Rose, N. S., McDaniel, M. A., & Kliegel, M. (2010). The age prospective memory paradox: young adults may not give their best outside of the lab. *Developmental Psychology, 46*(6), 1444-1453.

Bonin, P., Méot, A., Aubert, L.-F., Malardier, N., Niedenthal, P. M., & Capelle-Toczek, M.-C. (2003). Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots. *L'année Psychologique*, *103*(4), 655-694.

Brandimonte, M. A., Einstein, G. O., & McDaniel, M. A. (Eds.). (2014). *Prospective memory: Theory and applications*. New York and London: Psychology Press.

Castel, A. D., Middlebrooks, C. D., & McGillivray, S. (2016). Monitoring memory in old age: impaired, spared and aware. In Dunlosky, J., Tauber, S. U. K., & Tauber, P. D. R. S. (Eds.). (2016). *The Oxford Handbook of Metamemory* (pp. 519-535). Oxford University Press.

Cherry, K. E., & LeCompte, D. C. (1999). Age and individual differences influence prospective memory. *Psychology and Aging*, *14*(1), 60-76.

Devolder, P. A., Brigham, M. C., & Pressley, M. (1990). Memory performance awareness in younger and older adults, *Psychology and Aging, 5*(2), 291-303.

Einstein, G. O., & McDaniel, M. A. (1990). Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 717–726.

Einstein, G. O., & McDaniel, M. A. (1996). Retrieval processes in prospective memory: Theoretical approaches and some new findings. In M. Brandimonte, G. O. Einstein, & M. A. McDaniel (Eds.), *Prospective memory: Theory and applications* (pp. 115 – 141). New Jersey: Erlbaum.

Flavell, J. H. (1971). First discussant's comments: What is memory development the development of? *Human Development*, *14*, 272–278.

Gilbert, S. J. (2015). Strategic use of reminders: Influence of both domain-general and task-specific metacognitive confidence, independent of objective memory ability. *Consciousness and Cognition*, *33*, 245-260.

Henry, J. D., MacLeod, M. S., Phillips, L. H., & Crawford, J. R. (2004). A meta-analytic review of prospective memory and aging. *Psychology and Aging*, *19*(1), 27–39.

Hering, A., Phillips, L. H., & Kliegel, M. (2013). Importance effects on age differences in performance in event-based prospective memory. *Gerontology*, *60*(1), 73-78.

Ihle, A., Kliegel, M., Hering, A., Ballhausen, N., Lagner, P., Benusch, J., … Schnitzspahn, K. M. (2014). Adult age differences in prospective memory in the laboratory: are they related to higher stress levels in the elderly? *Frontiers in Human Neuroscience, 8*, 1021-1030.

Ihle, A., Schnitzspahn, K., Rendell, P. G., Luong, C., & Kliegel, M. (2012). Age benefits in everyday prospective memory: The influence of personal task importance, use of reminders and everyday stress. *Aging, Neuropsychology, and Cognition*, *19*(1-2), 84-101.

Kliegel, M., Mackinlay, R., & Jäger, T. (2008). Complex prospective memory: development across the lifespan and the role of task interruption. *Developmental psychology*, *44*(2), 612.

Knight, R. G., Harnett, M., & Titov, N. (2005). The effects of traumatic brain injury on the predicted and actual performance of a test of prospective remembering. *Brain Injury*, *19*(1), 19-27.

Kvavilashvili, L., & Ellis, J. (1996). Varieties of intention: Some distinctions and classifications. *Prospective memory: Theory and applications*, *6*, 183-207.

Meeks, J. T., Hicks, J. L., & Marsh, R. L. (2007). Metacognitive awareness of event-based prospective memory. *Consciousness and Cognition*, *16*(4), 997-1004.

Meier, B., Wartburg, P. Von, Matter, S., Rothen, N., & Reber, R. (2011). Performance predictions improve prospective memory and influence retrieval experience. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *65*(1), 12-18.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109-133.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*, 267–270.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, *26*, 125–322.

Phillips, L.H., Henry, J.D. Henry, Martin, M. (2008). Adult aging and prospective memory: The importance of ecological validity. In M. Kliegel, M.A. McDaniel, G.O. Einstein (Eds.), *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives* (pp. 161–185). Mahwah, NJ : Lawrence Erlbaum.

Rendell, P. G., & Craik, F. I. (2000). Virtual week and actual week: Age-related differences in prospective memory. *Applied Cognitive Psychology*, *14*(7), 43-62.

Rendell, P. G., & Thomson, D. M. (1999). Aging and prospective memory: Differences between naturalistic and laboratory tasks. *The Journals of Gerontology: Psychological Sciences and Social Sciences*, *54B*(4), 256-269.

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In Dunlosky, J., Tauber, S. U. K., & Tauber, P. D. R. S. (Eds.). (2016). *The Oxford Handbook of Metamemory* (pp. 519-535). *Oxford University Press.*

Rummel, J., Kuhlmann, B. G., & Touron, D. R. (2013). Performance predictions affect attentional processes of event-based prospective memory. *Consciousness and Cognition*, *22*(3), 729-741.

Schnitzspahn, K. M., Ihle, A., Henry, J. D., Rendell, P. G., & Kliegel, M. (2011a). The age-prospective memory-paradox: An exploration of possible mechanisms. *International Psychogeriatrics*, *23*(04), 583-592.

Schnitzspahn, K. M., & Kliegel, M. (2009). Age effects in prospective memory performance within older adults: The paradoxical impact of implementation intentions. *European Journal of Ageing*, *6*(2), 147-155.

Schnitzspahn, K. M., Zeintl, M., Jäger, T., & Kliegel, M. (2011b). Metacognition in prospective memory: are performance predictions accurate? *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *65*(1), 19-26.

Smith, R. E., & Bayen, U. J. (2006). The source of adult age differences in event-based prospective memory: A multinomial modeling approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 623.

Weaver III, C. A., Terrell, J. T., Krug, K. S., & Kelemen, W. L. (2008). The Delayed JOL Effect with very long delays: Evidence from flashbulb memories. *A handbook of memory and metacognition*, 155-172.

West, R. L. (1988). Prospective memory and aging. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current re- search and issues: Vol. 2. Clinical and educational implications* (pp. 119-128). Chichester, UK: Wiley.

Woods, S. P., Weinborn, M., Velnoweth, A., Rooney, A., & Bucks, R. S. (2012). Memory for intentions is uniquely associated with instrumental activities of daily living in healthy older adults. *Journal of the International Neuropsychological Society*, *18*(1), 134-138.

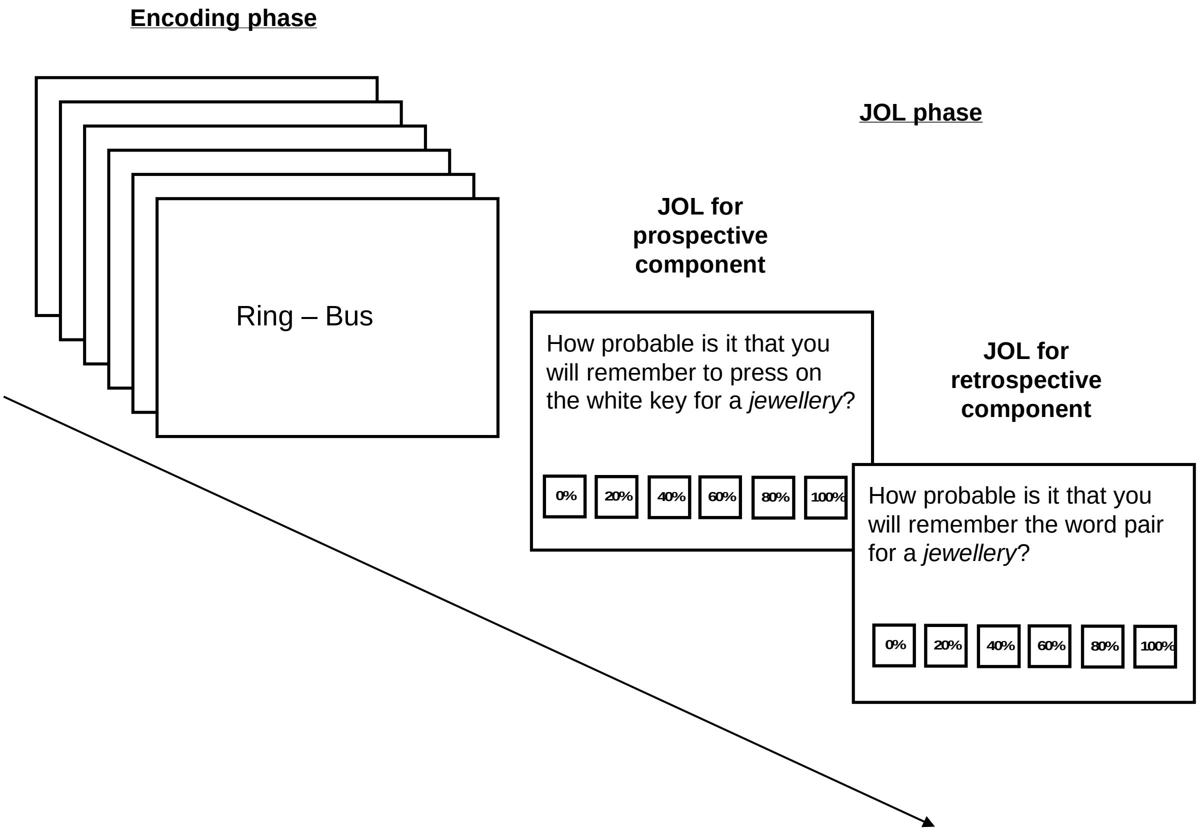
Encoding phase
Ring – Bus
JOL phase
JOL for prospective component
How probable is it that you will remember to press on the white key for a *jewellery*?
0%
20%
40%
60%
80%
100%
JOL for retrospective component
How probable is it that you will remember the word pair for a *jewellery*?
0%
20%
40%
60%
80%
100%

Figure 1. Encoding phase followed by the JOL phase with an example of a JOL for the word pair “Ring - Bus using the supra-ordinate category of the PM cue (i.e. jewellery).

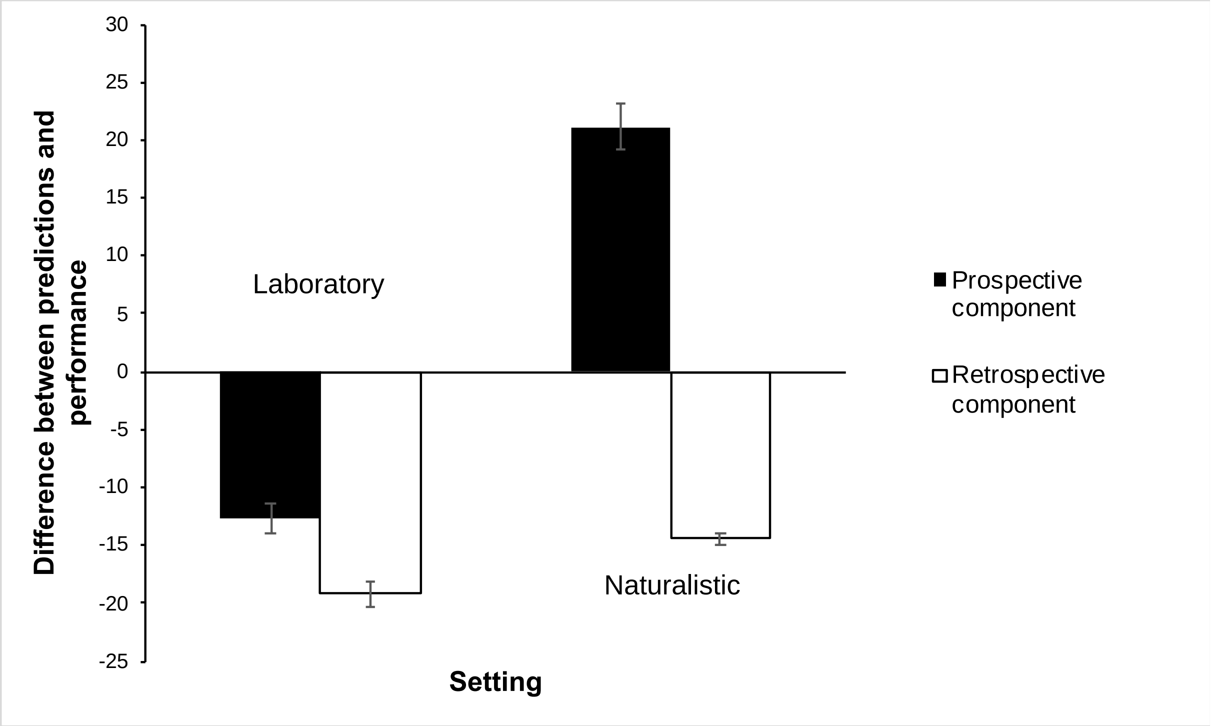

Figure 2. Differences between the predictions and the performance scores for the two components of the tasks in the two settings. Differences were conducted by subtracting the mean performance in % from the mean JOL in %. Thus, negative values represent participants' underconfidence and positive values their overconfidence.

Table 1. Descriptive (means and standard errors) and inferential statistics comparing the control and prediction groups on PM performance in the laboratory and naturalistic setting for the prospective and retrospective components

	Descriptive statistics		Inferential statistics	
	Prediction group *M (SE)*	Control group *M (SE)*	*t*(86)	*p*
Laboratory setting				
Prospective component	62.50 (27.93)	64.5 (23.71)	-0.365	0.716
Retrospective component	60.56 (37.02)	55.37 (33.70)	0.696	0.488
Naturalistic setting				
Prospective component	33.93 (24.59)	31.89 (29.71)	0.357	0.722
Retrospective component	78.27 (25.86)	81.39 (28.15)	-0.551	0.583

Table 2. Participants' mean scores and standard deviations for prospective and retrospective components of PM performance and predictions in the laboratory and naturalistic setting

	Laboratory		Naturalistic	
	Prospective	Retrospective	Prospective	Retrospective
Predictions				
M *(SD)*	52.41 (*2.79*)	44.03 (*3.37*)	55.90 (*2.28*)	66.65 (*3.31*)
Performance				
M *(SD)*	65.07 (3.*77*)	63.19 (*5.24*)	34.78 (*3.63*)	81.06 (*3.30*)

Note. All data are proportion correct.