Monitoring accuracy suffers when working memory demands increase: evidence of a dependent relationship

Donna Bryce, Florian Kattner, Paul Wellingerhof, Teresa Birngruber

Angaben zur Veröffentlichung / Publication details:

Bryce, Donna, Florian Kattner, Paul Wellingerhof, and Teresa Birngruber. 2023. "Monitoring accuracy suffers when working memory demands increase: evidence of a dependent relationship." *Journal of Experimental Psychology: Learning, Memory and Cognition* 49 (12): 1909–22. https://doi.org/10.1037/xlm0001262.

Monitoring Accuracy Suffers When Working Memory Demands Increase: Evidence of a Dependent Relationship

Donna Bryce[1, 2], Florian Kattner[3], Teresa Birngruber[1], and Paul Wellingerhof[1]
[1] Department of Psychology, Eberhard Karls University of Tübingen
[2] Department of Psychology, University of Augsburg
[3] Department of Psychology, Health and Medical University

Knowing what one knows and accurately monitoring one's own capacities and performance on a moment-to-moment basis are important determinants of task success. Individual differences in such metacognitive monitoring are well documented, but what determines an individual's monitoring accuracy in a particular context is yet to be fully understood. One candidate contributor to monitoring accuracy is working memory. In this study, we investigated whether and how working memory contributes to the accuracy of monitoring processes. Most evidence for a positive relationship between working memory and monitoring accuracy has been provided by correlational studies. Here, an experimental approach was applied in which confidence judgments were collected after each memory recall in three working memory experiments, and the effect of increasing the working memory demands on monitoring accuracy was examined. A visuospatial complex span task, a verbal complex span task, and an updating task served as the working memory tasks, to cover the range of methods used in working memory research. Confirmatory analyses conducted using cumulative link mixed models indicated that in two out of three experiments, monitoring accuracy suffered when working memory demands increased. As such, the weight of evidence supports a dependent relationship between working memory and monitoring processes, whereby monitoring accuracy can fluctuate during a task depending on the available cognitive resources. This indicates that the sensitivity of metacognitive monitoring is at least partly determined by the nature of the cognitive processing taking place in the primary task.

Keywords: metacognition, monitoring, working memory, updating, perceptual load

Knowing what one knows and accurately monitoring one's own capacities and performance on a moment-to-moment basis are important determinants of both immediate task success (Mengelkamp & Bannert, 2010) and educational achievement (Veenman et al., 2004). Individual differences in such metacognitive monitoring are well documented (Maki, 1998), but what determines an individual's monitoring accuracy in a particular context is yet to be fully understood. The current study aims to elucidate the cognitive basis of monitoring processes, with a particular focus on the competing or supporting role of working memory.

Donna Bryce https://orcid.org/0000-0001-8311-4457

We thank Meryem Banabak for data collection. This study was funded by the Program for the Promotion of Junior Researchers, University of Tübingen. The authors have no competing interests to declare.

All procedures performed were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments. Informed consent was obtained from all participants included in the study.

This study was not preregistered. Stimuli, data, and data simulation code are freely accessible on the Open Science Framework: https://osf.io/ar5ht/.

Correspondence concerning this article should be addressed to Donna Bryce, Department of Psychology, University of Augsburg, Universitätsstrasse 10, 86159 Augsburg, Germany. Email: donna.bryce@uni-a.de

Empirically, monitoring processes have been studied by examining their product, namely monitoring judgments provided by participants. Different aspects of the subjective experience are studied using different types of monitoring judgments, such as feeling of knowing, judgments of learning, and confidence judgments (CJs). These methods emerged from the field of metamemory research (e.g., Hart, 1967), and the judgments vary according to when they are collected in relation to a memory task (or any other primary task). In the current study, we focus on CJs that are collected after a memory task; these are judgments regarding how confident a participant is about the accuracy of the recall they just provided. The sensitivity of monitoring judgments to variations across trials is named relative monitoring accuracy and is classically assessed via correlational techniques (i.e., the gamma correlation between monitoring judgments and task performance) or signal detection theory analysis in the case of binary (correct/incorrect) outcomes (e.g., Maniscalco & Lau, 2012). Mengelkamp and Bannert (2010) reported that the relative accuracies of monitoring judgments collected at different time points in a comprehension task were not stable, but those collected in different tests both taken at the end of a task were stable. Gaining a better understanding of what determines variations in monitoring accuracy, both within and between individuals, is of considerable value as it will enable us to develop more successful methods to improve monitoring accuracy and maximize the accuracy of information gained through monitoring processes. One candidate for a determinant or contributor to monitoring accuracy is executive functions. In the developmental literature, executive

functions have even been proposed as precursors of metacognitive skills (Bryce et al., 2015; Roebers, 2017). Here we investigate in adults whether one executive function, the updating of working memory, contributes to the accuracy of monitoring processes.

Working memory is the type of memory required when some information must be manipulated and stored in memory for a short time, for instance, while remembering the names of items and simultaneously reordering them. Efficient updating of working memory is thought to be crucial for flexible responding, problem solving, and reasoning, and as such is considered one of three component executive functions (Miyake et al., 2000). Already 20 years ago a theoretical link was made between metacognitive skills (of which monitoring is one) and executive functions in a review by Fernandez-Duque et al. (2000). They stated that executive functions "may be the building blocks that metacognitively sophisticated thinkers use in their achievement of complex tasks" (p. 291). Since then, a wealth of empirical evidence has supported this link between executive functions and metacognitive monitoring, a selection of which is reviewed here.

Studies that provided evidence of a positive relationship between working memory and monitoring accuracy can be categorized into two types. The first are correlational studies, in which working memory and monitoring are assessed in separate tasks and a positive relationship is observed. For instance, Perrotin et al. (2007, 2008) observed that healthy participants with higher working memory capacity also tend to monitor their performance in another task more accurately, compared to those with lower working memory capacity. Griffin et al. (2008) observed that monitoring accuracy in a text comprehension task was positively related to participants' working memory capacity, but that rereading the text eliminated this relationship. They interpreted this finding as indicating that participants with lower working memory capacity were limited in their ability to encode metalevel cues during reading, but when the working memory demands were reduced by rereading, they could more successfully engage in metacognitive monitoring. Finally, Boduroglu et al. (2014) also reported a significant correlation between working memory span and CJs, but considered this an artifact that was driven by specific task design features. Indeed, these correlational studies are all likely to suffer from task-specific effects and fall short of providing insight regarding causality or dependency in the relationship between working memory and metacognitive monitoring processes.

The second type of study that provides evidence for a positive relationship between working memory and monitoring accuracy can be labeled "monitoring of memory" as monitoring is measured within the memory task itself. Here, monitoring judgments about performance within memory tasks are provided either after every trial (local judgments) or at the end of the task (global judgments). Some such studies examined whether the objective data pattern is reflected in monitoring judgments (e.g., Rademaker et al., 2012), whereas others examined whether monitoring accuracy varies between participant groups (younger vs. older adults, or those with low vs. high working memory capacity). For instance, Touron et al. (2010) observed better relative monitoring accuracy about one's own performance in a working memory task for younger than older adult participants, and Komori (2016) reported higher monitoring accuracy in participants with a high working memory capacity than those with a low working memory capacity. Consistent with this, Adam and Vogel (2017) observed a positive correlation between monitoring accuracy and overall working memory performance (both assessed in the same task) and speculated that this may emerge because individuals with better monitoring accuracy more promptly notice when their attention begins to drift and are able to redirect it to the task.

Some "monitoring of memory" studies additionally assessed the effect of manipulating working memory demands on monitoring accuracy. Schwartz (2008) examined "tip of the tongue" and "feeling of knowing" states and found that tip of the tongue states decreased (and feeling of knowing states increased or were unchanged) when working memory load increased. He speculated that the processes underlying working memory interfere with the metacognitive processes that lead to tip-of-the-tongue experiences. That is, when resources are used up by memory, insufficient resources remain for metacognitive processes (see Stine-Morrow et al., 2006, for a similar argument). A similar effect was observed in the Touron et al. (2010) study, whereby monitoring was more accurate for short than long sequence lengths (i.e., low vs. high working memory demands). These authors suggested that participants who monitor their performance more accurately than implement more successful strategies and therefore improve their memory performance. While this explanation may be intuitive from a between-subjects perspective, it is hard to reconcile this account with the within-subjects effect of better monitoring accuracy when working memory demands vary within one experiment. Indeed, these authors acknowledged that they could not exclude the alternative explanation, namely that those with better working memory can allocate more cognitive resources to monitoring processes.

In summary, while there is evidence of a positive relationship between working memory skills and monitoring accuracy, different post hoc explanations have been proposed for this relationship. Broadly speaking, these can be grouped and termed competition for resources, monitoring-based strategy improvements, and selection of valid cues. *Competition for resources* refers to the idea that monitoring processes either require working memory or the same cognitive resources as working memory and that when these resources are occupied maintaining items in working memory, insufficient resources remain for monitoring processes. This stems from direct access or trace-access view of monitoring (e.g., Nelson & Narens, 1990) whereby monitoring involves continuous online observation of our own task processing. According to this view, individuals with superior working memory capacity also monitor more successfully because maintaining items in working memory occupies fewer cognitive resources for them, and thus more resources can be directed toward efficient direct monitoring. Likewise, within one individual, we would hypothesize that when resources are used up by high working memory demands during a particular task, fewer resources remain for monitoring and thus monitoring accuracy suffers.

The proposition that *monitoring-based strategy improvements* are responsible for the relationship between working memory and monitoring accuracy stems from the idea of monitoring-based control (Koriat et al., 2006). As mentioned above, it has been argued that participants who monitor more accurately are more sensitive to fluctuations in their performance and therefore implement better strategies, thus improving their performance in working memory tasks. However, we would not expect that unpredictable variations in working memory demands during a task would lead to differences in monitoring accuracy, as the assumed direction of causation is the reverse.

Some other authors suggested a third explanation—that individuals with better executive functions are better able to *select and integrate valid cues* when generating their monitoring judgments, and this drives the positive relationship between the two skills (Perrotin et al., 2008). This account stems from a cue utilization or inferential view of monitoring (Koriat & Levy-Sadot, 2001), whereby various sources of information other than the directly accessed memory strength can contribute to monitoring judgments and the validity of these cues determines monitoring accuracy. Again, such an account can well explain between-subjects effects, but would not predict that variations in working memory demands during a task would lead to differences in monitoring accuracy. That is not to say that the selection and integration of cues is not resource demanding, but that the advantage afforded to individuals with higher working memory capacities takes effect at the point of generating the monitoring judgment. That is, resource depletion during a working memory trial would not affect cue selection and integration at the end of a trial. To date, few studies have aimed to find direct evidence regarding the reason(s) that working memory and monitoring accuracy are positively correlated. Here we seek evidence for or against these three explanations, by examining how monitoring accuracy is affected by changing working memory demands within a task.

The Current Study

In the current study, we attempt to overcome the interpretation challenges inherent in correlational studies by taking an experimental approach. As such, we collect CJs after each memory recall in working memory experiments and examine the effect of increasing the working memory demands on monitoring accuracy. We attempt to improve upon existing experimental approaches by manipulating not just working memory demands but also some other aspects of task difficulty (see below). This serves to establish whether any increase in task difficulty affects monitoring accuracy or whether there is a specific relationship between working memory and monitoring processes. If working memory and metacognitive monitoring are dependent on each other or make demands on the same cognitive resources, we would expect poorer monitoring accuracy when memory demands increase. However, if working memory and monitoring are positively correlated because individuals with better monitoring skills make more sensitive strategy adjustments in a working memory task, or because they more successfully select valid cues when producing a monitoring judgment, we would not expect working memory demands within a trial to affect monitoring accuracy.

Additionally, because working memory has been measured in a multitude of ways (Conway et al., 2005; Schmiedek et al., 2009), to ensure the generalizability of our findings, we applied this approach to three different types of working memory tasks. In Experiment 1, the primary task was a complex span task tapping the visuospatial domain. In this "odd one out" task (adapted from Alloway, 2007), participants must identify a deviant shape and remember its location. In Experiment 2, the primary task was a verbal complex span task in which participants had to process the meaning of a sentence and remember the final word of the sentence (adapted from Experiment 2 of Daneman & Carpenter, 1980). In Experiment 3, the primary task was an updating task where participants had to continuously update the contents of their working memory to retain the final four or six items presented (adapted from Dahlin et al., 2008). The working memory demands were increased by increasing the number of to-be-recalled items (Experiments 1–3), and task difficulty was manipulated via visual similarity (Experiment 1), semantic complexity (Experiment 2), and visual degradation (Experiment 3). Figure 1 provides an illustration of one trial of each experiment, and more details are provided in each Method section.

Experiment 1: Visuospatial Complex Span Task

Experiment 1 was a visuospatial complex span task in which participants had to identify a series of deviant shapes and then recall the location of these shapes. Working memory demands were manipulated via the number of to-be-recalled locations (six or eight), and to examine the impact of an additional (perceptual) task difficulty manipulation, the visual similarity of the to-be-compared shapes was manipulated. Specifically, visual similarity was increased in half of the trials to make perceptual discrimination of the deviant shape from the stimulus set more challenging.

Method

Participants

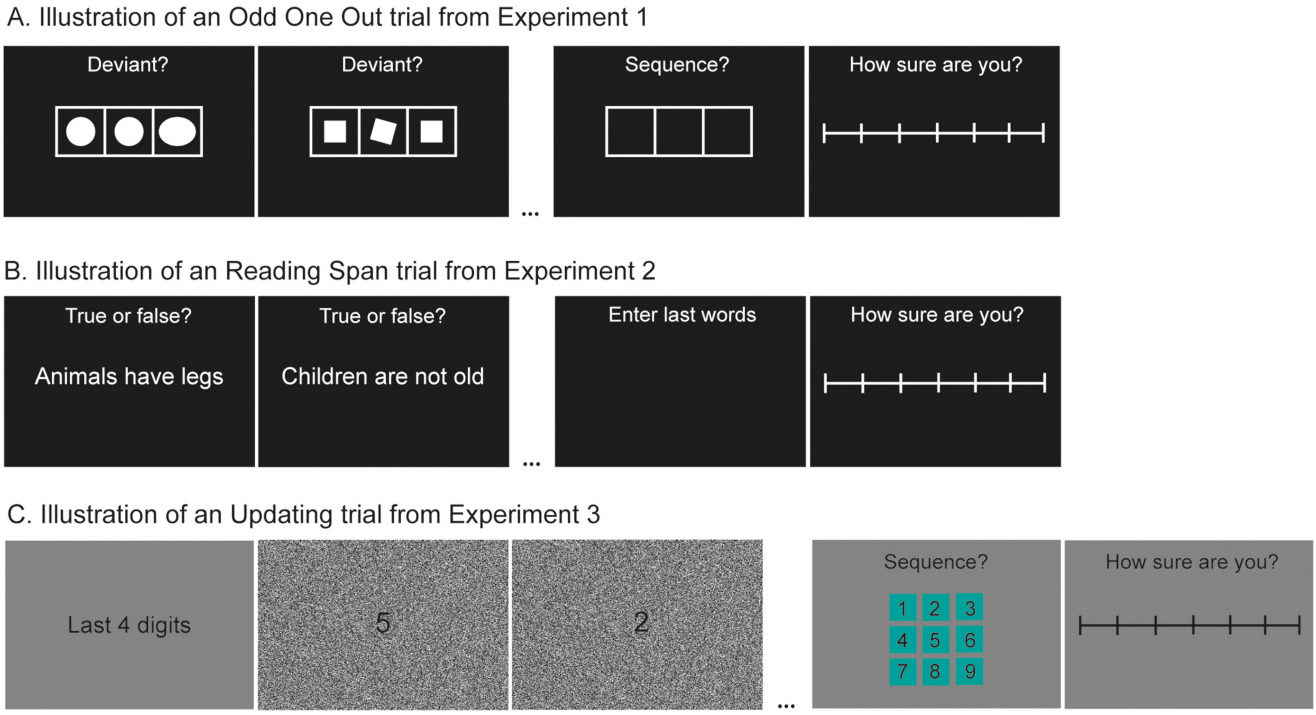
For each experiment reported here, a sample size of 20 was selected. This sample size is large enough to observe the most relevant effect sizes reported in the literature to date, namely those from Touron et al. (2010). In that study, monitoring accuracy (the gamma correlation) was higher in trials with short than long sequence lengths, and the related Cohen's d values were 0.89 (Experiment 1) and 0.68 (Experiment 2). A sensitivity analysis for a paired-samples t-test, computed using G*Power (Faul et al., 2007), indicated a sample size of 20 provides 90% power to observe effects of at least $d = 0.68$. Accordingly, 20 participants with a mean age of 20.6 years (range 18–29) participated in Experiment 1 for course credit at the University of Tübingen. Eighteen were right handed (two were left handed), three identified as male (17 as female), and all had normal or corrected vision. All procedures performed were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments. Informed consent was obtained from all participants included in the study.

Apparatus and Stimuli

This and all subsequent experiments were conducted on a Mac computer via MATLAB with the PsychToolbox extension (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997), in a sound-attenuated testing booth. Participants gave responses via mouse click.

For the entirety of an odd one out trial, three square frames arranged horizontally were presented in white in the center of a black screen. To make an "Easy" odd one out stimulus, one of these frames was filled in white. To make a "Hard" odd one out stimulus, a geometric shape was presented in white within each frame. For each of these trios, two shapes were identical and one deviated slightly from the others. The basic geometric shapes used included a circle, two types of crosses (+ and ×), a square, a diamond, a triangle, and a circle intersected by a thick line. To make the "odd" deviant shape, one of the shapes was slightly modified. These modifications included increasing or decreasing the size of the shape, rotating the shape, or slightly altering the dimensions of the shape

Figure 1
Illustration of One Trial of Experiments 1–3



Note. In the odd one out task, participants identified the deviant in a trio of geometric shapes, for a sequence of six or eight trios, reported the locations of the deviants in sequence, and then provided a confidence judgment regarding their recall accuracy (A). In the reading span task, participants judged the veracity of sentences, for a series of four or six sentences, reported the last word of each sentence in sequence, and then provided a confidence judgment regarding their recall accuracy (B). In the updating task, participants were informed by a cue how many ("*n*") digits they would be required to recall, and then were presented a series of digits. Following a variable updating duration, they recalled the "*n*" last digits they had stored in memory and gave a confidence judgment regarding their recall accuracy (C). See the online article for the color version of this figure.

(e.g., presenting an oval with two circles). A set of 132 such "hard" odd one out trios were generated in a pilot study and can be found online (https://osf.io/ar5ht/). This pilot study with $n = 10$ participants indicated that participants were 510 ms slower, $t(9) = -15.33$, $p < .001$, and 2.84% less accurate, $t(9) = 7.87$, $p < .001$, whenidentifying the deviant in "hard" than in "easy" odd one out trios.

CJs were collected on a 7-point Likert scale. This was a white horizontal line with seven small vertical tick marks presented on black background. The left-most tick mark was labeled "sehr unsicher" (*very unsure*), the third tick mark was labeled "ziemlich unsicher" (*quite unsure*), the fifth tick mark was labeled "ziemlich sicher" (*quite sure*), and the seventh right-most tick mark was labeled "sehr sicher" (*very sure*).

Procedure and Design

An illustration of an odd one out trial is provided in Figure 1A. Each trial began with the presentation of the three empty frames for 500 ms followed by an odd one out trio and the prompt "Abweichler?" (*deviant?*). The trio remained on the screen until the participant clicked with the mouse within one frame to indicate their selection of the deviant shape. Next, the empty frames were presented for another 500 ms, followed by a new trio of shapes, and so on until a sequence of either six or eight trios had been presented. The sequence lengths of six and eight were selected based on pilot testing with the aim to avoid ceiling and floor effects. After making a series of six or eight odd one out judgments, the empty white frames were presented again on the screen with the prompt "Reihenfolge?" (*sequence?*) and participants were instructed to click in the frames to indicate the locations in which the deviants had appeared, in the same order in which they had appeared. Participants were explicitly informed that if they made an error in identifying the odd one out of a trio, they should report the correct location of the odd one out in the sequence recall, rather than the erroneously reported location. When a frame was selected, it turned red for 200 ms. When a sequence of six (or eight) trios had been presented, a sequence of six (or eight) reported locations was required before proceeding to the CJ. The Likert scale was presented on the screen with the prompt "Wie sicher bist du?" (*how sure are you?*) and participants clicked with the mouse on the scale to give one of seven discrete CJ values regarding their confidence in their sequence recall. A small round white marker was presented for 1 s on the Likert scale to indicate the value selected by the participant. A blank black screen was presented for 500 ms before the next trial began.

The experiment had a 2 (sequence length: six, eight) × 2 (task difficulty: easy, hard) within-subjects design, resulting in four unique trial types, which each repeated 28 times throughout the experimental session. In total, 392 hard and 392 easy odd one out trios were presented ($28 \times 6 + 28 \times 8$). When a hard trio was presented in the experiment, one of the 132 trios was sampled randomly without

replacement until all had been presented once; then, this procedure was repeated twice more. On average, each of the 132 hard trios was presented 3 times throughout the experiment. For each trio within each sequence, the location of the deviant was randomly chosen. Participants first completed eight practice trials, followed by 14 experimental blocks composed of eight trials each. Within each block of eight trials, each unique trial type was repeated twice and was presented in a random order. Participants could take self-paced breaks between the blocks, and the whole session lasted approximately 75 min.

Data Analysis

Data were analyzed using R (R Core Team, 2021). Practice trials were discarded before analysis. Four dependent measures were analyzed via repeated measures analyses of variance with the within-subject factors sequence length (2) and task difficulty (2). The four dependent measures were: the mean time taken to identify deviants in a trial, the proportion of correctly identified deviants in a trial, the proportion of the sequence correctly recalled, and mean CJs. For the recall accuracy score, one point was awarded for each item correctly recalled in the correct serial position in a sequence and this value was converted to a proportion by dividing by the sequence length (six or eight). For ANOVAs, the Greenhouse–Geisser correction was used to adjust p-values where appropriate, and partial eta-squared (η_p^2) effect sizes are provided.

Metacognitive monitoring accuracy and the influence of working memory and task difficulty demands on monitoring accuracy were assessed via cumulative link mixed models (CLMMs). Predictor variables were recall accuracy (centered), sequence length, and task difficulty (both categorical variables were treatment coded). This approach to analysis has the advantage of assessing what contributes to CJs on a single trial level, and CLMMs were employed rather than linear mixed effect models due to the ordinal nature of the dependent variable (CJs were given on a 7-point Likert scale). Monitoring accuracy is defined as the relationship between monitoring judgments and performance; as such, a significant effect of recall accuracy within the CLMM indicates monitoring accuracy (and is the equivalent of a significant correlation in typical resolution measures). An interaction of another factor with recall accuracy, for example, a Sequence Length × Recall Accuracy interaction, indicates that the extent to which recall accuracy contributes to CJs is affected by this factor (e.g., sequence length). This can be interpreted as indicating that monitoring accuracy is affected by sequence length. The CLMM analysis aimed to determine the best fitting model via model selection (following the procedure described by Barr et al., 2013). The full model included all predictors and their two-way interactions as fixed effects. The model was iteratively reduced by one term and compared to the more complex model using likelihood ratio tests. The best fitting CLMM is presented in tabular form. CLMMs were conducted using the R package *ordinal* (Christensen, 2019) using a logit link function, and random intercepts were permitted to vary for each participant.

In the spirit of a multiverse analysis (Steegen et al., 2016), we also analyzed our data in the more traditional way of first calculating gamma correlations between recall accuracy and CJ for each participant and each experimental condition and submitting these to a repeated measures ANOVA with the within-subjects factors sequence length and task difficulty. Goodman–Kruskal gamma correlations were calculated using the R package *RoCoCo* (Bodenhofer et al., 2013; Bodenhofer & Klawonn, 2008).

An additional correlational analysis analogous to the typical individual differences approach taken in other studies was conducted with these data. First, each participant's dataset was split in two (odd-numbered trials comprised one dataset, even-numbered trials comprised the other). With one half of the data, the gamma correlation between CJ and recall accuracy was calculated as a measure of relative monitoring accuracy. With the other half of the data, mean recall accuracy was calculated as a measure of working memory performance. As a measure of the relationship between working memory performance and monitoring accuracy, a Pearson correlation between the two variables was calculated for the sample.

Stimuli and data for each of the three experiments are freely accessible on the Open Science Framework: https://osf.io/ar5ht/.

Results

Working Memory Performance and Confidence Judgments

On average, it took participants 1,134 ms to identify a deviant in an easy task difficulty trial, and 1,846 ms to identify a deviant in a hard task difficulty trial, and this was reflected in a significant main effect of task difficulty, $F(1, 19) = 154.38$, $p < .001$, $\eta_p^2 = .89$. Time to identify a deviant was not affected by sequence length, $F(1, 19) = 0.03$, $p = .855$, $\eta_p^2 < .01$, nor was there a significant Sequence Length × Task Difficulty interaction, $F(1, 19) = 0.66$, $p = .427$, $\eta_p^2 = .03$. Similarly, participants correctly identified a higher proportion of deviants in easy ($M = 0.98$) than hard ($M = 0.93$) trials, $F(1, 19) = 41.24$, $p < .001$, $\eta_p^2 = .68$, and identification accuracy did not vary as a function of sequence length, $F(1, 19) = 0.28$, $p = .605$, $\eta_p^2 = .01$. However, a significant interaction, $F(1, 19) = 4.76$, $p = .042$, $\eta_p^2 = .20$, indicated that the task difficulty effect on the proportion of correctly identified deviants was slightly larger when sequence length was eight than when it was six.

Recall accuracy can be seen in Figure 2A. Participants correctly recalled a greater proportion of the sequence when it was easy to identify the deviants than when it was hard, $F(1, 19) = 29.54$, $p < .001$, $\eta_p^2 = .61$, and when sequence length was six than when it was eight, $F(1, 19) = 72.66$, $p < .001$, $\eta_p^2 = .79$. The two factors did not significantly interact, $F(1, 19) = 0.02$, $p = .887$, $\eta_p^2 < .01$.

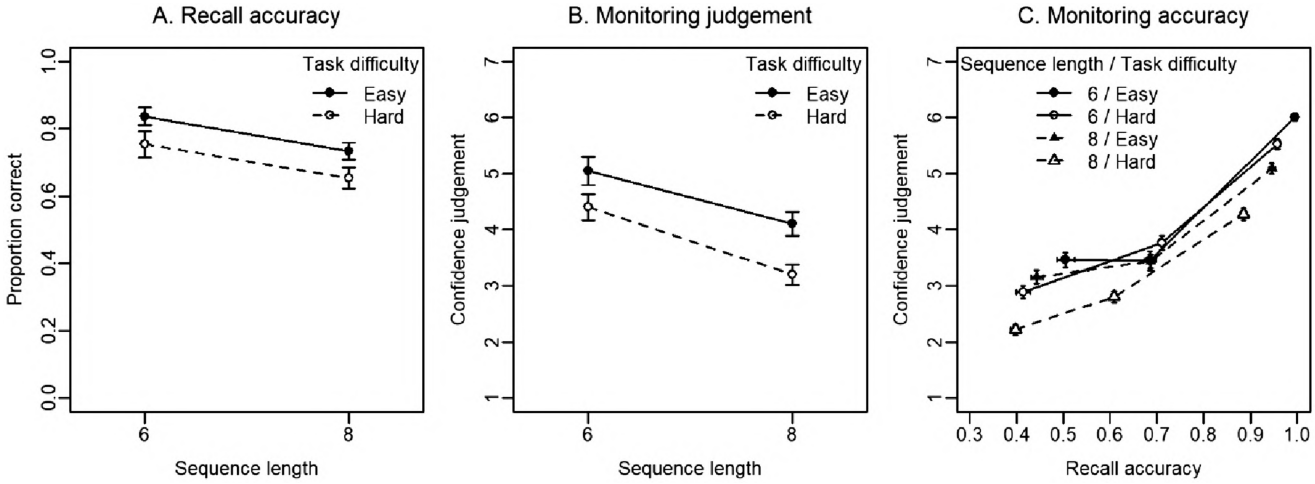
As can be seen in the mean CJs displayed in Figure 2B, participants reported higher confidence when task difficulty was easy than hard, $F(1, 19) = 39.43$, $p < .001$, $\eta_p^2 = .67$, and when sequence length was six than when it was eight, $F(1, 19) = 82.50$, $p < .001$, $\eta_p^2 = .81$. There was no significant interaction, $F(1, 19) = 1.91$, $p = .183$, $\eta_p^2 = .09$.

Metacognitive Monitoring Accuracy

The best fitting CLMM is presented in Table 1 and for illustrative purposes vincentized data of the relationship between recall accuracy and CJ is presented in Figure 2C. Likelihood ratio tests indicated that recall accuracy significantly improved the model, $\chi^2(1) = 869.96$, $p < .001$, and the positive estimate in Table 1 reveals a positive association between participants' CJs and recall accuracy, that is, significant monitoring accuracy. The best fitting CLMM additionally indicated a significant effect of sequence length, $\chi^2(1) = 105.86$, $p < .001$, a significant effect of task difficulty, $\chi^2(1) = 59.00$, $p < .001$ as well as significant interactions of

Figure 2

Results From the Visuo-Spatial Complex Span Task in Experiment 1



Note. Mean recall accuracy (A), monitoring judgment (B), and monitoring accuracy (C) as a function of sequence length and task difficulty. In panels A and B, trials composed of easy-to-identify deviants are depicted in solid lines, trials containing hard-to-identify deviants are depicted in dashed lines, error bars represent ± 1 *SE*. In panel C, confidence judgment is plotted against recall accuracy. For illustration single trials were vincentized based on recall accuracy (single trial data entered the data analysis). That is, the first vincentile in each panel represents 33% of trials with the lowest recall accuracy for each participant by condition; the second contains the next 33% of trials with lowest recall accuracy, and so on. Error bars represent ± 1 within-subject *SE*.

sequence length and recall accuracy, $\chi^2(1) = 6.41$, $p = .011$, and of sequence length and task difficulty, $\chi^2(1) = 5.75$, $p = .016$. The crucial interaction for our hypothesis is the Sequence Length × Recall Accuracy interaction and it is clear from Figure 2C and the negative estimate value in Table 1 that this reflects that objective performance (recall accuracy) contributes less to the monitoring judgment as the sequence length increases (solid lines are slightly steeper than dashed lines). This could be interpreted as poorer monitoring accuracy as working memory demands increase. The Sequence Length × Task Difficulty interaction perhaps requires some further explanation—this shows that the contribution task difficulty makes to CJs decreases as sequence length increases.

The repeated measures ANOVA of gamma correlations (a more traditional measure of the relative monitoring accuracy) confirmed the above data pattern. Crucial for our hypothesis, gamma correlations were higher in the low ($\gamma = 0.71$) than high ($\gamma = 0.59$) sequence length,[1] $F(1, 19) = 8.24$, $p = .010$, $\eta_p^2 = .30$. Additionally, a significant Sequence Length × Task Difficulty interaction reveals a larger effect of sequence length on monitoring accuracy when the task was easy than hard, $F(1, 19) = 4.51$, $p = .047$, $\eta_p^2 = .19$. A main effect of task difficulty was not observed, $F(1, 19) = 1.03$, $p = .323$, $\eta_p^2 = .05$.

Table 1

Estimates for Fixed Effects in the Best Fitting CLMM for Experiment 1

Predictor	Estimate (*SE*)	95% CI	*p* Value
Recall acc.	0.06 (0.003)	[0.06, 0.07]	<.001
Task diff.[a]	−0.40 (0.11)	[−0.62, −0.18]	<.001
Seq. length[b]	−0.60 (0.11)	[−0.82, −0.38]	<.001
Recall Acc. × Seq. Length	−0.83 (0.32)	[−0.01, −0.002]	.009
Task Diff. × Seq. Length	−0.41 (0.16)	[−0.72, −0.11]	.008

Note. Confidence judgment is the dependent variable, *p*-values are based on the Wald statistic. CLMM = cumulative link mixed model; CI = confidence interval.
[a] Baseline: easy. [b] Baseline: sequence length of six.

Individual Differences Analysis

The Pearson correlation between individual participants' monitoring accuracy score (gamma correlation coefficient calculated with half of the trials) and their working memory task performance (mean recall accuracy calculated with the other half of the trials) was $r = .77$, $p < .001$. This correlation can be interpreted as indicating that individuals who perform better in the odd one out task also monitor their own performance in the odd one out task more accurately.

Discussion

Both manipulations (visual similarity and sequence length) affected working memory performance—recall was worse when the task was harder and when more items had to be recalled from working memory. Participants had rather accurate introspections about their own performance, as their mean CJs reflected the same data pattern. In terms of sensitivity to trial-by-trial variations in performance, participants showed rather accurate monitoring, which slightly suffered when participants had to maintain longer sequences in memory. This data pattern is consistent with the idea that holding items in working memory and metacognitive monitoring have a dependent relationship. One possible interpretation is that when

[1] To enable comparison with the effect sizes used in the power analysis, the data were also analyzed with a paired-samples *t*-test and Cohen's *d* calculated. There was a significant effect of sequence length, $t(19) = 3.20$, $p = .005$, $d = 0.72$.

cognitive resources were occupied processing the primary working memory task, fewer resources were left for monitoring processes and monitoring accuracy suffered. Specifically, working memory load affected monitoring accuracy, whereas visual similarity (i.e., perceptual task load) did not. In Experiment 2, we applied the same experimental approach to a different type of working memory task, that is, one involving sentence processing, to evaluate the generalizability of this finding.

Experiment 2: Verbal Complex Span Task

Experiment 2 was a verbal complex span task, in which participants had to judge the veracity of sentences and then recall the last word of these sentences. Working memory demands were manipulated via the number of to-be-recalled words (four or six), and to examine the impact of another task difficulty manipulation the semantic complexity of the sentences was manipulated in three ways (see Apparatus and Stimuli below). The sentences were designed such that a deeper level of processing would be required to correctly judge the veracity in more complex sentences.

Method

Participants

Twenty participants (14 identified as female, six as male) recruited from the University of Tübingen completed the experiment for payment or course credit. Their mean age was 24.0 years (range 20–36), all were right handed and all reported normal or corrected vision.

Apparatus and Stimuli

In Experiment 2, participants gave responses to the working memory task via computer keyboard and gave their CJs via mouse. Sentences that were to be judged as true or false were presented in white, Arial font, size 18 in the center of a black screen. Moreover, 372 sentences were generated: 186 designed to be easy and 186 designed to be hard. The difficulty of sentences was manipulated in three different ways: contrast distance, negation, and category exemplar. Contrast distance refers to whether a comparison is made between two units that are far apart (e.g., *a day is longer than a minute*; easy) or close together (e.g., *an hour is longer than a minute*; hard). Negation (e.g., Dudschig & Kaup, 2020) refers to the fact that confirmations (e.g., *Easter is in spring*; easy) are easier to process than negations (e.g., *Easter is not in winter*; hard). Category exemplar refers to the hierarchical nature of categories (e.g., Vandierendonck, 1991), whereby categories at higher levels of the hierarchy include categories of lower levels. We created easy sentences by transcending only one level of the hierarchy (e.g., *animals can move*) and hard sentences by transcending more than one level (e.g., *lizards can move*). Sentences were created in sets of four within which two were true and two were false, two were easy and two were hard, and either all four final words were identical or two sets of identical final words were identical. The word length of the sentences ranged from three to seven, with a median of four words. Word length was also kept mostly constant within the sets of four. The full set of sentences (in the German language) is available online (https://osf.io/ar5ht/). A pilot study (with $n = 10$) indicated that participants were 292 ms slower, $t(9) = -6.44$, $p < .001$, and 3.92% less accurate, $t(9) = 2.70$, $p = .024$, to correctly respond to the hard than the easy sentences.

CJs were collected on the 7-point Likert scale described previously for Experiment 1. Participants gave their judgments via mouse click.

Procedure and Design

An illustration of a Reading Span trial is provided in Figure 1B. Each trial began with the presentation of a sentence, randomly selected from the easy or hard sentences depending on the trial type. Participants pressed the "J" button (for *Ja*, yes) to indicate that the sentence was true, and the "F" button (for *Falsch*, false) to indicate that the sentence was false. Participants were explicitly instructed to respond based on whether the sentence was *typically* true or false. After responding to a sentence, a 500-ms blank screen was presented before the next sentence. In this experiment, sequence lengths of four and six were selected after pilot testing was conducted to achieve a recall accuracy of approximately 70%. After a sequence of four or six sentences had been presented, participants were asked to type the last word of each sentence in the correct order, prompted by the message "Please enter the last word of each sentence and press Enter after each word (1/4)." After entering all four or six words, a 500-ms blank screen followed and then the participants indicated their confidence in their sequence recall on the 7-point Likert scale by clicking with the mouse. The intertrial interval was 1 s. In the practice block, if participants did not give their veracity judgment within 4 s, the sentence was removed and they received immediate feedback that they should respond more quickly. In the experimental blocks, if participants did not give their veracity judgment within 3 s the trial continued and they received feedback at the end of the trial.

The experiment had a 2 (sequence length: four, six) × 2 (task difficulty: easy, hard) within-subjects design, resulting in four unique trial types, which each repeated 15 times throughout the experimental session. In total, 150 hard and 150 easy sentences were presented ($15 \times 4 + 15 \times 6$); they were randomly selected without replacement from the 186 sentences in each difficulty set. Participants first completed four practice trials, followed by six experimental blocks composed of 10 trials each. Participants could take self-paced breaks between the blocks, and the whole session lasted approximately 60 min.

Data Analysis

One participant provided a CJ of one on 59 out of 60 trials. Their data were not included in further analysis. Practice trials were not included in analysis. As in Experiment 1, there were four dependent measures: the mean response time to give a veracity judgment in a trial, the proportion of correct veracity judgments in a trial, the proportion of the sequence correctly recalled (recall accuracy), and the mean CJ. In calculating the recall accuracy, one point was awarded for each correct word in the correct position in a sequence and this value was converted to a proportion depending on whether it was a four- or six-item sequence. Words with minor typographical errors were accepted as correct. Each dependent variable listed above was entered into a repeated measures analysis of variance with the within-subject factors sequence length (2) and task difficulty (2). The Greenhouse–Geisser correction was used to adjust p-values where appropriate, and partial eta-squared (η_p^2) effect sizes are

provided. The monitoring accuracy and individual differences analysis was conducted in the same way as for Experiment 1.

Results

Working Memory Performance and Confidence Judgments

Participants were 209 ms faster to respond to the sentences in easy than hard trials, $F(1, 18) = 99.77$, $p < .001$, $\eta_p^2 = .85$, and a significant interaction indicated that this task difficulty effect was larger when sequence length was four than when it was six, $F(1, 18) = 7.46$, $p = .014$, $\eta_p^2 = .29$. There was, however, no main effect of sequence length on mean response time to give veracity judgments, $F(1, 18) = 0.18$, $p = .678$, $\eta_p^2 = .01$. Participants also made more accurate veracity judgments in easy (0.89) than hard trials (0.81), $F(1, 18) = 70.19$, $p < .001$, $\eta_p^2 = .80$, and in short (0.86) than long (0.84) sequences, $F(1, 18) = 4.74$, $p = .043$, $\eta_p^2 = .21$. The interaction of task difficulty and sequence length was not significant, $F(1, 18) = 0.26$, $p = .617$, $\eta_p^2 = .01$.

As can be seen in Figure 3A, participants' memory recall was better when the sentences had been easy than hard, $F(1, 18) = 7.20$, $p = .015$, $\eta_p^2 = .29$, and in short than long sequences, $F(1, 18) = 126.05$, $p < .001$, $\eta_p^2 = .88$, indicating that both manipulations affected working memory performance. The two factors did not, however, interact, $F(1, 18) = 0.05$, $p = .823$, $\eta_p^2 < .01$.

Figure 3B depicts mean monitoring judgments, and it can be seen that participants reported higher confidence regarding their recall accuracy in easy than hard trials, $F(1, 18) = 5.95$, $p = .025$, $\eta_p^2 = .25$, and in short than long sequences, $F(1, 18) = 107.48$, $p < .001$, $\eta_p^2 = .86$. As with the performance data, the two factors did not interact, $F(1, 18) = 2.76$, $p = .114$, $\eta_p^2 = .13$.

Metacognitive Monitoring Accuracy

Participants in Experiment 2 also showed a positive association between recall accuracy and their CJs; in other words, accurate monitoring. The best fitting CLMM (summarized in Table 2) indicated significant effects of recall accuracy, $\chi^2(1) = 363.83$, $p < .001$ and sequence length, χ^2 $(1) = 38.19$, $p < .001$, only. These effects reflect, on the one hand, a general sensitivity to trial-by-trial variation in performance (i.e., monitoring accuracy), and on the other hand, that as sequence length increases, lower CJs are given (dashed lines shifted downward compared to solid lines in Figure 3C). Important for our hypothesis, the Sequence Length × Recall Accuracy interaction did not significantly improve the model, reflected in the observation from Figure 3C that solid lines and dashed lines have a very similar slope.

The repeated measures ANOVA of gamma correlations confirmed the above data pattern. Gamma correlations were high ($\gamma = .57$) and unaffected by sequence length,[2] $F(1, 18) = 0.70$, $p = .415$, $\eta_p^2 = .04$, task difficulty, $F(1, 18) = 2.84$, $p = .109$, $\eta_p^2 = .14$, and there was no significant Sequence Length × Task Difficulty interaction, $F(1, 18) < 1$, $p = .968$, $\eta_p^2 < .01$.

Individual Differences Analysis

The Pearson correlation between individual participants' monitoring accuracy score (gamma correlation coefficient calculated with one half of trials) and their working memory task performance (mean recall accuracy calculated with the other half of trials) was $r = .58$, $p = .009$. In contrast to the equivalent result from Experiment 1, this result indicates a moderate between-subjects relationship between recall accuracy and monitoring accuracy.

Discussion

As in Experiment 1, both the semantic complexity and sequence length manipulations affected recall accuracy in the reading span task. Again, on the mean level, the objective data pattern was well reflected in participants' CJs, indicating that participants were aware of the impact of these manipulations on their performance, and the CLMM confirmed a significant contribution of recall accuracy to CJs (i.e., high relative monitoring accuracy). Unlike Experiment 1, monitoring accuracy was not affected by any of the experimental manipulations. Instead, only performance on the task and the sequence length of to-be-recalled words contributed to CJs. One possible explanation for this is that the semantic complexity manipulation utilized in this experiment operated very differently than the visual similarity manipulation utilized in Experiment 1. This will be elaborated on in the General Discussion. Next, in Experiment 3, a different type of working memory task structure was employed, namely an updating design rather than a complex span design.

Experiment 3: Updating Task

Experiment 3 was an updating task, in which participants had to continuously update the contents of their working memory with a certain number of previously presented digits. Working memory demands were manipulated via the number of to-be-recalled digits (four or eight). To examine the impact of another task difficulty manipulation, in half of the trials, the digits were visually degraded to increase perceptual task-encoding load.

Method

Participants

Twenty participants with a mean age of 24.5 years (range 20–36) were recruited from the University of Tübingen and took part in the study for course credit or payment. Five identified as male and 15 as female, three were left handed (17 were right handed), and all reported normal or corrected vision.

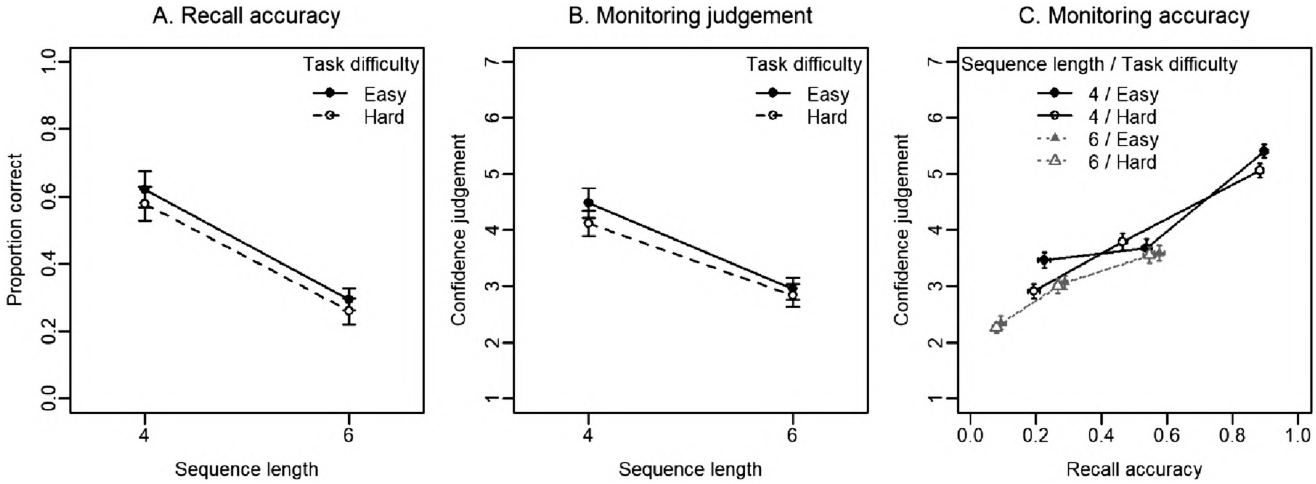
Apparatus and Stimuli

Stimuli were 284 × 256-pixel images of black digits from 1 to 9 on a white background. Static Gaussian noise of either low (Var = 1) or high variance (Var = 10) was added to the image (using the function "imnoise" in MATLAB and a similar procedure as in Kattner & Bryce, 2022) to manipulate perceptual task difficulty. To the participants, the background screen appeared gray throughout the experiment. Participants provided their serial recall via mouse click on a numeric pad presented on the screen. CJs were given as in the previous experiments on a 7-point Likert scale, this time presented in black on a gray background.

[2] To enable comparison with the effect sizes used in the power analysis, the data were also analyzed with a paired-samples *t*-test and Cohen's *d* calculated. There was no significant effect of sequence length, $t(18) = 1.12$, $p = .277$, $d = 0.26$.

Figure 3

Results From the Verbal Complex Span Task in Experiment 2



Note. Mean recall accuracy (A), monitoring judgment (B), and monitoring accuracy (C) as a function of sequence length and task difficulty. In panels A and B, trials composed of sentences whose veracity was easy to judge are depicted in solid lines, trials containing sentences whose veracity was hard to judge are depicted in dashed lines, error bars represent ± 1 *SE*. In panel C, confidence judgment is plotted against recall accuracy. For illustration, single trials were vincentized based on recall accuracy (single trial data entered the data analysis). That is, the first vincentile in each panel represents 33% of trials with the lowest recall accuracy for each participant by condition; the second contains the next 33% of trials with lowest recall accuracy, and so on. Error bars represent ± 1 within-subject *SE*.

Procedure and Design

An illustration of one trial is provided in Figure 1C. Each trial began with a cue, presented for 1 s, informing participants how many (*n*) digits they should maintain in memory and report at the end of the sequence (e.g., "Letzte 4 Ziffern"; *last four digits*). This was followed by a 200-ms blank screen, and then a sequence of digits was presented in succession each for 1 s. These digits could be easy or hard to identify, manipulated via perceptual noise. At the end of the sequence of digits, there was a 200-ms interval before the numeric pad was presented and participants entered the "*n*" last digits in the order they recalled seeing them. After another 200 ms blank screen, the CJ was collected via mouse click. An intertrial interval of 1 s then followed.

Two levels of *n* (four and six) were selected in this experiment based on pilot testing aiming for a recall accuracy of approximately 70%. Before being prompted to recall the "*n*" last digits, participants experienced a sequence ("run") of digits that varied in length. Sequences could be $n + 0$, 1, 2, 3, 4, 5, or 6, translating to run lengths of 4–10 and 6–12 (when *n* is four and six, respectively). It was deemed important to include some run lengths of $n + 0$ to ensure that participants really engaged in updating from the start of the trial. In run lengths of nine or less, unique digits were randomly selected and presented in random order. In run lengths of 10 or more, direct repetitions of digits within the run were prohibited.

The experiment had a 2 (*n*: four, six) × 2 (task difficulty: easy, hard) within-subjects design, resulting in four unique trial types, which each repeated 40 times throughout the experimental session. Participants first completed 12 practice trials, followed by 10 experimental blocks composed of 16 trials each. Participants could take self-paced breaks between the blocks, and the whole session lasted approximately 90 min.

Table 2

Estimates for Fixed Effects in the Best Fitting CLMM for Experiment 2

Predictor	Estimate (*SE*)	95% CI	*p* Value
Recall acc.	0.04 (0.002)	[0.04, 0.05]	<.001
Seq. length [a]	−0.79 (0.13)	[−1.04, −0.54]	<.001

Note. Confidence judgment is the dependent variable, *p*-values are based on the Wald statistic. CLMM = cumulative link mixed model; CI = confidence interval.

[a] Baseline: sequence length of four.

Data Analysis

Performance in the 12 practice trials was not analyzed. Recall accuracy, that is, the proportion of the *n* digits correctly recalled, and CJs were analyzed in repeated measures analysis of variance with the within-subjects factors *n* (2) and task difficulty (2). The Greenhouse–Geisser correction was used to adjust *p*-values where appropriate, and partial eta-squared effect (η_p^2) sizes are provided. As in Experiments 1 and 2, a CLMM analysis tested for the best fitting model from a full model containing the effects of *n*, task difficulty and recall accuracy and their two-way interactions. The individual differences analysis was conducted following the method described in Experiment 1.

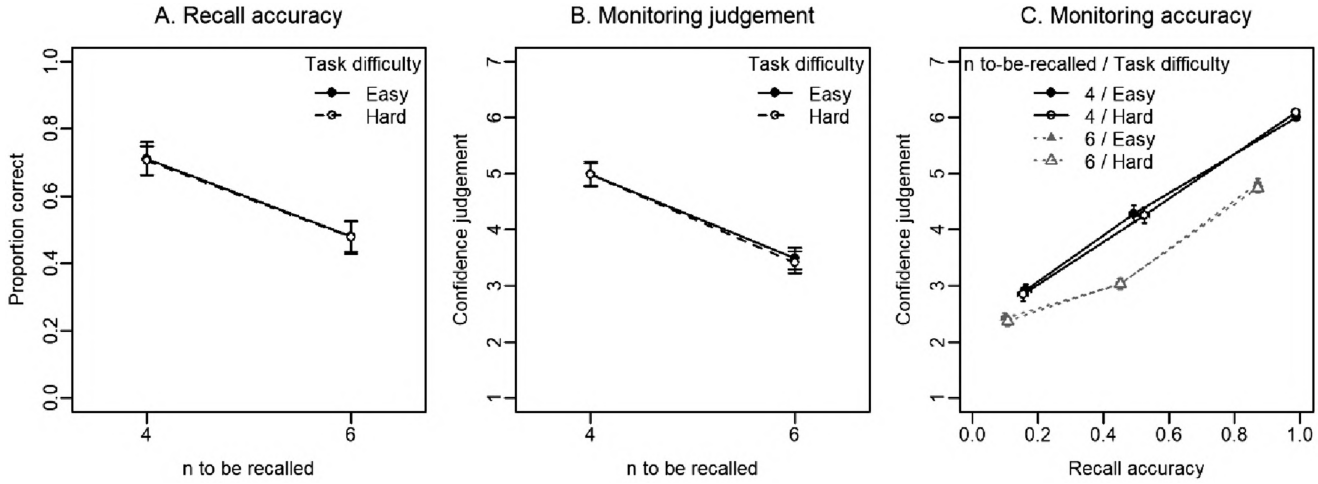
Results

Working Memory Performance and Confidence Judgments

Memory recall is depicted in Figure 4A. Participants recalled a higher proportion of items in the "recall 4" than "recall 6" condition, $F(1, 19) = 83.48$, $p < .001$, $\eta_p^2 = .81$, whereas task difficulty (i.e., the level of perceptual noise added to digits) did not affect recall, $F(1, 19) = 0.15$, $p = .702$, $\eta_p^2 = .01$ and the two-way interaction did not reach significance, $F(1, 19) = 0.03$, $p = .867$, $\eta_p^2 < .01$.

Figure 4

Results From the Updating Task in Experiment 3



Note. Mean recall accuracy (A), monitoring judgment (B), and monitoring accuracy (C) as a function of *n* to-be-recalled and task difficulty. In panels A and B, trials with low perceptual noise are depicted in solid lines, trials with high perceptual noise are depicted in dashed lines, error bars represent ± 1 *SE*. In panel C, confidence judgment is plotted against recall accuracy. For illustration, single trials were vincentized based on recall accuracy (single trial data entered the data analysis). That is, the first vincentile in each panel represents 33% of trials with the lowest recall accuracy for each participant by condition; the second contains the next 33% of trials with lowest recall accuracy, and so on. Error bars represent ± 1 within-subject *SE*.

The same effects were reflected in CJs (Figure 4B). While participants reported higher confidence in the "recall 4" than the "recall 6" condition, $F(1, 19) = 122.69$, $p < .001$, $\eta_p^2 = .87$, task difficulty did not affect CJs, $F(1, 19) = 0.84$, $p = .371$, $\eta_p^2 = .04$, and the interaction did not reach significance, $F(1, 19) = 0.47$, $p = .503$, $\eta_p^2 = .02$.

Metacognitive Monitoring Accuracy

As in previous experiments, a positive relationship between recall accuracy and CJs can be observed in Figure 4C, reflecting monitoring accuracy. The best fitting CLMM, reported in Table 3, contained recall accuracy, $\chi^2(1) = 1{,}724.60$, $p < .001$ and *n* to-be-recalled, $\chi^2(1) = 228.23$, $p < .001$, as well as the significant interaction of *n* to-be-recalled and recall accuracy, $\chi^2(1) = 39.86$, $p < .001$. This interaction indicates that increasing the working memory demand in this experiment resulted in recall accuracy contributing less to CJs, which can be interpreted as poorer monitoring accuracy when participants had to recall more digits. This pattern can be observed in Figure 4C (solid black lines steeper than gray dashed lines).

The repeated measures ANOVA of gamma correlations confirmed the above data pattern. Gamma correlations were higher when participants had to recall fewer ($\gamma = 0.81$) than more ($\gamma = 0.63$) digits,[3] $F(1, 19) = 61.04$, $p < .001$, $\eta_p^2 = .76$. There was no significant effect of task difficulty, $F(1, 19) = 0.66$, $p = .427$, $\eta_p^2 = .03$, nor was there a significant Sequence Length × Task Difficulty interaction, $F(1, 19) = 0.17$, $p = .638$, $\eta_p^2 = .01$.

Individual Differences Analysis

The Pearson correlation between individual participants' monitoring accuracy score (gamma correlation coefficient calculated with half of the trials) and their working memory task performance (mean recall accuracy calculated with the other half of the trials) was extremely high in Experiment 3, $r = .92$, $p < .001$. Similar to Experiment 1, this indicates that the better one performs in this updating task, the more accurately they can monitor their own performance.

Discussion

Overall, these results were consistent with the results of Experiment 1—increasing the working memory demands resulted in less sensitive trial-by-trial monitoring. This is consistent with the idea that when cognitive resources are occupied by the primary working memory updating task, fewer resources are available to effectively monitor and consequently monitoring judgments become less accurate. In this experiment, manipulating the task difficulty via visual degradation did not affect memory recall, CJs or monitoring accuracy. One could question whether this manipulation was successful in terms of increasing perceptual task-encoding load. Interestingly, the individual differences analyses also indicated a very high correlation between participants' performance on the updating task and their monitoring accuracy, which is also consistent with Experiment 1 and previous literature.

General Discussion

In a series of metamemory experiments, we aimed to better understand the role that working memory plays in monitoring processes.

[3] To enable comparison with the effect sizes used in the power analysis, the data were also analyzed with a paired-samples *t*-test and Cohen's *d* calculated. There was a significant effect of sequence length, $t(19) = 6.78$, $p < .001$, $d = 1.51$.

Table 3
Estimates for Fixed Effects in the Best Fitting CLMM for Experiment 3

Predictor	Estimate (*SE*)	95% CI	*p* Value
Recall acc.	0.05 (0.002)	[0.05, 0.06]	<.001
n to-be-recalled[a]	−1.06 (0.07)	[−1.20, −0.93]	<.001
Recall Acc. × *n*	−0.01 (0.002)	[−0.01, −0.007]	<.001

Note. Confidence judgment is the dependent variable, *p*-values are based on the Wald statistic. CLMM = cumulative link mixed model; CI = confidence interval.
[a] Baseline: *n* to-be-recalled of four.

More specifically, we aimed to establish whether working memory and monitoring processes have a dependent relationship or whether a positive relationship between the two skills is observed for other reasons. To this end, across three experiments working memory demands and some other aspects of task difficulty were varied, and the impact of these manipulations on relative monitoring accuracy was assessed. Due to the heterogeneity of methods used in working memory research, this approach was applied to three different types of working memory tasks: a visuospatial complex span task, a verbal complex span task, and an updating task. Likewise, task difficulty was manipulated in different ways in each experiment, namely via visual similarity, semantic complexity, and visual degradation. In summary, in all three experiments, participants demonstrated high levels of relative monitoring accuracy, as recall accuracy (objective performance) made a large and significant contribution to CJs (monitoring judgment). Further, in Experiments 1 and 3, and consistent with Touron et al. (2010), monitoring accuracy was negatively impacted by increases in working memory demands, delivering evidence for a dependent relationship between working memory and monitoring processes. As such, the weight of evidence from this series of experiments supports a *competition for resources* explanation for the relationship between working memory and monitoring. Alternative explanations, challenges to this conclusion and open questions are discussed in the following.

As reviewed in the Introduction, three explanations for the positive relationship between working memory and monitoring accuracy have been outlined in the literature. Aside from the *competition for resources* account, these include the idea that individuals with better monitoring skills go on to implement more effective working memory strategies (*monitoring-based strategy improvements*) and that individuals with better working memory can better select and integrate cues when giving their monitoring judgments (*selection of valid cues*). As already discussed in the Introduction, it is hard to reconcile the *monitoring-based strategy improvements* explanation with the within-participants effects reported here. Indeed, such monitoring-based strategy improvements may exist and contribute to the correlations reported in the individual differences analyses (namely, that better monitoring accuracy in one half of the experiment is associated with better working memory performance in the other half of the experiment). However, it is hard to imagine that differences in monitoring accuracy across conditions drive the differences in working memory performance within one experiment, particularly when participants are not able to predict the working memory demands of the upcoming trial. Indeed, as acknowledged by Griffin et al. (2008), "it is unclear how any construct that could reasonably be labeled metacognitive monitoring would not be a secondary process to whatever is being monitoring [sic] at the object level" (p. 100). That is, we posit that while this explanation may hold some truth for the observation that individuals with better monitoring also perform better in working memory tasks, it is not sufficient to explain the effects observed here.

Evidence for the use of cues to produce monitoring judgments, and thereby for the latter explanation, can be evaluated based on the data collected here. That is, if information other than objective performance contributes to CJs, this would be reflected in a significant effect of that variable in the best fitting CLMM and a vertical shift in the lines in panel C of each results figure. For instance, in Experiment 1, we can see that when sequence length was eight (gray lines), task difficulty made a greater contribution to CJs than when sequence length was six. One could interpret this as showing that when working memory was maximally taxed, participants used task difficulty as a cue and thus lowered their CJ by a relatively fixed amount when the task was hard. This does not reflect accurate relative monitoring, as the slope of the Recall Accuracy × Confidence Judgment function does not change. In Experiment 2, we observe a significant effect of sequence length in the CLMM, indicating that longer sequences receive lower CJs over and above what would be expected based on objective performance. As such, we could interpret this as indicating that sequence length was used as a cue. In Experiment 3, we also observe that the number of items to be recalled makes a significant contribution to CJs, over and above what would be expected based on objective performance. As such, data from all three experiments indicate that participants utilized various cues in making their monitoring judgments. That is, as well as using the most valid information available to them (namely, recall accuracy) participants used other contextual features of the trial when providing their monitoring judgments (namely, sequence length and task difficulty). However, the strong contribution of objective performance to CJs in all three experiments suggests that participants do indeed engage in a direct "online" type of monitoring in this context. The fact that relative monitoring accuracy was affected by fluctuations in working memory demands across trials in two out of three experiments provides strong evidence for a dependent relationship between them, even if monitoring judgments are additionally influenced by other sources of information.

Recent contributions from the field of computational neuroscience have drawn similar conclusions as those above. The primary aim of these studies has been to establish what information forms the basis of confidence (or uncertainty) judgments in visual working memory[4] tasks, and the evidence is consistent with the idea that participants can engage in direct access monitoring of the strength of their memory representations (Geurts et al., 2022; Honig et al., 2020; Li et al., 2021; van den Berg et al., 2017). More specifically, an fMRI-derived measure of the encoding noise associated with a specific memory item is positively related to both the accuracy and variability of memory recall, and the confidence reported by participants about their recall (Honig et al., 2020; Li et al., 2021). Importantly, these authors also highlight that a lot of unexplained noise affects CJs (Geurts et al., 2022), that individuals' prior expectations regarding stimuli can also contribute to metacognitive judgments (Honig et al., 2020), and that other sources of information

[4] Note that while the authors of these studies refer to their tasks as assessing visual working memory, they notably differ from the ones employed in the current study as they do not require information manipulation.

such as the time taken to recall the memory also positively correlate with encoding noise and participants' CJs (Li et al., 2021). The findings from the current study suggest that not only are there likely interindividual differences in the information that forms the basis of confidence in our memory recall, but there may also be intraindividual differences as well. That is, the weighting of information may vary across experimental conditions, and across trials with different working memory demands. From this perspective, perhaps an increased working memory load hinders the system's ability to extract or evaluate the probabilistic information regarding the precision of the memory representation, resulting in other sources of information contributing more to our feeling of confidence.

The main challenge to the conclusion that working memory and monitoring are in a dependent relationship concerns the lack of support for this from Experiment 2. In this verbal complex span task, although monitoring was quite accurate, monitoring accuracy was not affected by variations in working memory demands. Another notable way in which Experiment 2 differed from the other experiments is in the overall working memory performance—on average participants recalled only 44% of the words in Experiment 2, compared to 74% of the deviant locations in Experiment 1 and 59% of digits in Experiment 3. However, there does not appear to be a floor effect at play here. CJs reflected these overall differences in performance, with average CJs of 3.6 in Experiment 2, and 4.2 in Experiments 1 and 3. Further, the task difficulty manipulation in Experiment 2 (increased semantic complexity of sentences) arguably necessitated a deeper level of processing than was required in Experiments 1 and 3 (where some aspect of perceptual load was manipulated). Thus, the processing of the sentences in Experiment 2 may have drawn on the same cognitive resources as working memory and monitoring. As such, perhaps Experiment 2 was so challenging that participants' resources were almost fully occupied processing the primary working memory task, making concurrent, direct monitoring (or the extraction of probabilistic information regarding encoding precision) no longer possible. Consequently, participants may have generated their monitoring judgments differently in this context, namely with a greater reliance on cues available posttask (such as their beliefs about the effect of sequence length on their memory). Thus, the within-task manipulations of working memory load did not further affect monitoring accuracy (see Griffin et al., 2008 for a similar argument). One can also infer from the current results that monitoring was overall less accurate in Experiment 2 than the other experiments, as the β-estimate for recall accuracy and the mean gamma correlations were much lower.

A strength of the current study is that the data could also be analyzed analogously to how such data have been analyzed in previous studies, namely an individual differences approach. In all three experiments, these analyses indicated that individuals who perform better in working memory tasks also monitor their own performance more accurately. This finding is consistent with Adam and Vogel (2017) and also with the interpretation that working memory and monitoring compete for resources. That is, individuals who have a greater working memory capacity and/or are more skilled at updating their working memory use up fewer cognitive resources when engaging in working memory tasks and therefore have more resources available to them to simultaneously monitor. The fact that the individual differences correlation was descriptively stronger in experiments in which the working memory demands also affected monitoring accuracy (Experiments 1 and 3) is consistent with the within- and between-participants effects having the same source (although it must be acknowledged that a formal test of the relatedness of these effects was not possible).

Limitations

The current study is not exempt from criticism and some limitations should be acknowledged. One challenge to the generalizability of these findings regards the method used to collect CJs. In all three experiments, a visual analog scale labeled "very unsure" to "very sure" was employed to collect CJs. This decision was taken to avoid providing participants with additional cues regarding the sequence length of the trial just processed and to maintain the same scale across conditions and experiments. However, it is not known whether these data patterns would be replicated with alternative ways of providing CJs (i.e., an absolute estimate of number of items correctly recalled). As has already been acknowledged, monitoring is not a unitary process (McDonough et al., 2021) and different effects can be observed when judgments are collected via different measures (Mengelkamp & Bannert, 2010). As such, testing the generalizability of these findings with other types of monitoring judgments may be a fruitful line of subsequent research. Further, our use of rather abstract labels for the scale may have introduced variability regarding how participants mapped their internal feeling of confidence to this scale. Individual differences in how people map their feeling of confidence to a scale are assumed, but also a challenge to evaluate within the current dataset. We have, however, conducted simulations to investigate if any systematic approaches to mapping confidence to the scale could account for our data pattern (in the case where there were no true differences in monitoring accuracy), and a report of this can be found at https://osf.io/ar5ht/. Based on these simulations, we consider it extremely unlikely that the data patterns observed in Experiments 1 and 3 are mere artifacts of our confidence reporting method.

The samples recruited are highly educated young adults and as such somewhat distinct from the previous studies reviewed in the Introduction section, which often focused on aging populations. One could question whether this sample is representative of the wider population and more recent advances in online experimentation may make it possible to recruit a sample that is more diverse in terms of age and education in future studies. While a range of tasks and task features were intentionally employed in order to assess the generalizability of the findings across different working memory tasks, this task heterogeneity could limit our ability to draw strong conclusions. In particular, the fact that overall performance and the nature of the task difficulty manipulation are confounded in Experiment 2 makes it hard to pinpoint why the results deviated in Experiment 2. Future studies could directly examine this verbal complex span task in more detail, for instance by comparing the effect of a perceptual load and cognitive load manipulation within the same experiment. While the current study achieved its aim of manipulating working memory load and observing the impact on monitoring accuracy, the selected experimental design had one drawback—sequence length and working memory load are confounded. An alternative design in which participants' working memories are taxed outside of the primary task (e.g., as a third task) may provide further clarity regarding the role of working memory in monitoring processes. On a theoretical level, one could question whether a strong association between performance and monitoring judgments is conclusive evidence that direct

access monitoring is taking place. It is conceivable that participants use some other cues that happen to be highly correlated with their performance. The approaches being applied in computational neuroscience reviewed above seem promising to address this issue. While of theoretical interest, understanding the basis of confidence may be less important than understanding what affects the accuracy of our confidence (although the two endeavors may prove inseparable). At least for self-regulation, it is important that participants gain accurate information through their monitoring processes, on which to base subsequent adjustments in their behavior.

Implications

The current findings suggest that in cases where monitoring accuracy is particularly important, we should make efforts to offload or reduce the working memory demands of the task to improve our chances of making accurate monitoring judgments. This raises the question whether working memory training to increase capacity or optimize updating would also lead to improvements in monitoring accuracy. In turn, an exciting avenue for future research would be to investigate whether we can flexibly allocate our cognitive resources between working memory encoding and/or maintenance and our monitoring processes, or whether the primary task always takes precedence. Further, we posit that these findings not only apply to this rather artificial context in which participants have to monitor their working memory performance directly, but that this interaction between working memory and monitoring would occur in any task with a working memory component. Indeed, these findings mirror recent findings in a dual-task processing context whereby monitoring accuracy was greatly improved when the working memory demands of the task were reduced (Bryce & Bratzke, 2022). One remaining open question is why are working memory and monitoring in a dependent relationship? Which cognitive resource are they in competition for, or (taking the computational neuroscience perspective) why would increased working memory demands reduce our access to a measure of encoding precision of a memory representation? In our view, further research is needed to definitively answer this, but the current results suggest new lines of research that could be followed to consider not only interindividual differences but also intra-individual differences in the information contributing to monitoring judgments. Further, these results highlight another mechanism by which working memory may contribute to learning and academic success. Not only does it assist with the processing of a primary task, but it may allow for more efficient metacognitive monitoring which in turn allows individuals to more effectively direct and control their own future learning.

Conclusion

The current study aimed to elucidate the cognitive basis of metacognitive monitoring, with a particular focus on the role of working memory. Previous correlational studies reported a positive relationship between performance on working memory tasks and the accuracy of monitoring processes. Taking an experimental approach, we directly tested the effect of increasing working memory demands on the accuracy of monitoring judgments and observed in two out of three experiments that monitoring accuracy suffered when working memory demands increased. These findings are consistent with a dependent relationship between monitoring accuracy and working memory. As such, monitoring accuracy can fluctuate during a task depending on the current working memory demands, and individuals with poorer working memories are less able to effectively monitor themselves simultaneously. These findings provide important insights into what might determine an individual's monitoring accuracy in a particular context, suggesting a dynamic interaction between the cognitive processing taking place in the primary task and metacognitive monitoring processes.

References

Adam, K. C. S., & Vogel, E. K. (2017). Confident failures: Lapses of working memory reveal a metacognitive blind spot. *Attention, Perception, & Psychophysics*, *79*(5), 1506–1523. https://doi.org/10.3758/s13414-017-1331-8

Alloway, T. P. (2007). *Automated working: Memory assessment: Manual*. Pearson.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bodenhofer, U., & Klawonn, F. (2008). Robust rank correlation coefficients on the basis of fuzzy orderings: Initial steps. *Mathware and Soft Computing*, *15*(1), 5–20.

Bodenhofer, U., Krone, M., & Klawonn, F. (2013). Testing noisy numerical data for monotonic association. *Information Sciences*, *245*, 21–37. https://doi.org/10.1016/j.ins.2012.11.026

Boduroglu, A., Tekcan, A. I., & Kapucu, A. (2014). The relationship between executive functions, episodic feeling-of-knowing and confidence judgements. *Journal of Cognitive Psychology*, *26*(3), 333–345. https://doi.org/10.1080/20445911.2014.891596

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Bryce, D., & Bratzke, D. (2022). The surprising role of stimulus modality in the dual-task introspective blind spot: A memory account. *Psychological Research*, *86*(4), 1332–1354. https://doi.org/10.1007/s00426-021-01545-y

Bryce, D., Whitebread, D., & Szücs, D. (2015). The relationships among executive functions, metacognitive skills and educational achievement in 5 and 7 year-old children. *Metacognition and Learning*, *10*(2), 181–198. https://doi.org/10.1007/s11409-014-9120-4

Christensen, R. H. B. (2019). *ordinal—Regression models for ordinal data*. R package (Version 2019.12-10) [Computer software]. https://CRAN.R-project.org/package=ordinal

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. *Science*, *320*(5882), 1510–1512. https://doi.org/10.1126/science.1155466

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Dudschig, C., & Kaup, B. (2020). Negation as conflict: Conflict adaptation following negating vertical spatial words. *Brain and Language*, *210*, Article 104842. https://doi.org/10.1016/j.bandl.2020.104842

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Executive attention and metacognitive regulation. *Consciousness and Cognition*, *9*(2), 288–307. https://doi.org/10.1006/ccog.2000.0447

Geurts, L. S., Cooke, J. R. H, van Bergen, R. S., & Jehee, J. F. M. (2022). Subjective confidence reflects representation of Bayesian probability in cortex. *Nature Human Behaviour*, *6*(2), 294–305. https://doi.org/10.1038/s41562-021-01247-w

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, *36*(1), 93–103. https://doi.org/10.3758/Mc.36.1.93

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning and Verbal Behavior*, *6*(5), 685–691. https://doi.org/10.1016/S0022-5371(67)80072-0

Honig, M., Ma, W. J., & Fougnie, D. (2020). Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. *Proceedings of the National Academy of Sciences*, *117*(15), 8391–8397. https://doi.org/10.1073/pnas.1918143117

Kattner, F., & Bryce, D. (2022). Attentional control and metacognitive monitoring of the effects of different types of task-irrelevant sound on serial recall. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(2), 139–158 https://doi.org/10.1037/xhp0000982

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in psychtoolbox-3? *Perception*, *36*(14), 1–16. https://doi.org/10.1068/v070821

Komori, M. (2016). Effects of working memory capacity on metacognitive monitoring: A study of group differences using a listening span test. *Frontiers in Psychology*, *7*, Article 285. https://doi.org/10.3389/fpsyg.2016.00285

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 34–53. https://doi.org/10.1037/0278-7393.27.1.34

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*(1), 36–69. https://doi.org/10.1037/0096-3445.135.1.36

Li, H.-H., Sprague, T. C., Yoo, A. H., Ma, W. J., & Curtis, C. E. (2021). Joint representation of working memory and uncertainty in human cortex. *Neuron*, *109*(22), 3699–3712. https://doi.org/10.1016/j.neuron.2021.08.022

Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 131–158). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410602350-13

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

McDonough, I. M., Enam, T., Kraemer, K. R., Eakin, D. K., & Kim, M. (2021). Is there more to metamemory? An argument for two specialized monitoring abilities. *Psychonomic Bulletin & Review*, *28*(5), 1657–1667. https://doi.org/10.3758/s13423-021-01930-z

Mengelkamp, C., & Bannert, M. (2010). Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. *Memory & Cognition*, *38*(4), 441–451. https://doi.org/10.3758/Mc.38.4.441

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). Academic Press. https://doi.org/10.1016/S0079-7421(08)60053-5

Pelli, D. G. (1997). The video toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897X00366

Perrotin, A., Belleville, S., & Isingrini, M. (2007). Metamemory monitoring in mild cognitive impairment: Evidence of a less accurate episodic feeling-of-knowing. *Neuropsychologia*, *45*(12), 2811–2826. https://doi.org/10.1016/j.neuropsychologia.2007.05.003

Perrotin, A., Tournelle, L., & Isingrini, M. (2008). Executive functioning and memory as potential mediators of the episodic feeling-of-knowing accuracy. *Brain and Cognition*, *67*(1), 76–87. https://doi.org/10.1016/j.bandc.2007.11.006

Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of Vision*, *12*(13), Article 21. https://doi.org/10.1167/12.13.21

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review*, *45*, 31–51. https://doi.org/10.1016/j.dr.2017.04.001

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1089–1096. https://doi.org/10.1037/a0015730

Schwartz, B. L. (2008). Working memory load differentially affects tip-of-the-tongue states and feeling-of-knowing judgments. *Memory & Cognition*, *36*(1), 9–19. https://doi.org/10.3758/MC.36.1.9

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Stine-Morrow, E. A. L., Shake, M. C., Miles, J. R., & Noh, S. R. (2006). Adult age differences in the effects of goals on self-regulated sentence processing. *Psychology and Aging*, *21*(4), 790–803. https://doi.org/10.1037/0882-7974.21.4.790

Touron, D. R., Oransky, N., Meier, M. E., & Hines, J. C. (2010). Metacognitive monitoring and strategic behaviour in working memory performance. *Quarterly Journal of Experimental Psychology*, *63*(8), 1533–1551. https://doi.org/10.1080/17470210903418937

van den Berg, R., Yoo, A. H., & Ma, W. J. (2017). Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychological Review*, *124*(2), 197–214. https://doi.org/10.1037/rev0000060

Vandierendonck, A. (1991). Primary and secondary generalization in categorization and typicality of well-defined concepts. *Bulletin of the Psychonomic Society*, *29*(6), Article 475. https://doi.org/10.1037/e665402011-031

Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction*, *14*(1), 89–109. https://doi.org/10.1016/j.learninstruc.2003.10.004