Atten Percept Psychophys (2017) 79:1506–1523
DOI 10.3758/s13414-017-1331-8

# Confident failures: Lapses of working memory reveal a metacognitive blind spot

**Kirsten C. S. Adam**[1,2] · **Edward K. Vogel**[1,2]

Published online: 3 May 2017


**Abstract** Working memory performance fluctuates dramatically from trial to trial. On many trials, performance is no better than chance. Here, we assessed participants' awareness of working memory failures. We used a whole-report visual working memory task to quantify both trial-by-trial performance and trial-by-trial subjective ratings of inattention to the task. In Experiment 1 ($N$ = 41), participants were probed for task-unrelated thoughts immediately following 20% of trials. In Experiment 2 ($N$ = 30), participants gave a rating of their attentional state following 25% of trials. Finally, in Experiments 3a ($N$ = 44) and 3b ($N$ = 34), participants reported confidence of every response using a simple mouse-click judgment. Attention-state ratings and off-task thoughts predicted the number of items correctly identified on each trial, replicating previous findings that subjective measures of attention state predict working memory performance. However, participants correctly identified failures on only around 28% of failure trials. Across experiments, participants' metacognitive judgments reliably predicted variation in working memory performance but consistently and severely underestimated the extent of failures. Further, individual differences in metacognitive accuracy correlated with overall working memory performance, suggesting that metacognitive monitoring may be key to working memory success.

**Electronic supplementary material** The online version of this article (doi:10.3758/s13414-017-1331-8) contains supplementary material, which is available to authorized users.

✉ Kirsten C. S. Adam
kadam1@uchicago.edu

[1] Department of Psychology, University of Chicago, 5848 S. University Ave., Chicago, IL 60637, USA

[2] Institute for Mind & Biology, University of Chicago, 940 E. 57th St, Chicago, IL 60637, USA

**Keywords** Visual working memory · Attentional control · Metacognition

Visual working memory is a highly limited memory system for temporarily representing a small amount of information from the environment. The measured capacity of this system differs substantially between individuals and is considered to be a stable trait of the observer that impacts performance on a wide variety of other cognitive tasks (e.g., Engle, Kane, & Tuholski, 1999; Unsworth, Fukuda, Awh, & Vogel, 2014). Recent work examining these individual differences has revealed that despite being highly stable between testing sessions, an individual's apparent capacity appears to fluctuate substantially from trial to trial within a testing session (Adam, Mance, Fukuda, & Vogel, 2015). These fluctuations in performance have been proposed to reflect variability in consistently engaging attentional control. When attentional control is fully engaged, subjects tend to reach a maximum capacity that is common across most individuals; when it is fully disengaged (e.g., an attentional lapse), working memory fails as subjects are often near chance performance. High-capacity individuals have far fewer of these fully or partially disengaged trials than low-capacity individuals, suggesting that these performance fluctuations within a session reveal an important determinant of individual differences in capacity.

What are the underlying causes of fluctuations in working memory performance, and are participants aware of their failures? While there is substantial evidence in the literature that individuals have access to reliable information about the contents of working memory (Fougnie, Suchow, & Alvarez, 2012; Mutluturk & Boduroglu, 2014; Rademaker, Tredway, & Tong, 2012; Vandenbroucke et al., 2014), there is reason to believe that they might systematically underestimate the frequency of working memory failures, especially if working

memory failures are related to being "off task." For example, observers severely underestimate the frequency of mind-wandering when monitoring their own performance relative to being "caught" by a computer-guided probe (Schooler, Reichle, & Halpern, 2004; Schooler et al., 2011). When observers are unaware of mind-wandering episodes, performance decrements are more severe (Smallwood, McSpadden, & Schooler, 2007). Consequently, subjects may not always be self-aware when they have completely disengaged from the task.

The present study is split into two distinct but related parts. First, in Experiments 1 and 2, we measured the covariation between task performance and subjective ratings of task-related and task-unrelated thoughts using a procedure in which subject's "thought contents" are probed on a random subset of trials. We predicted that working memory performance should broadly covary with subjective ratings of on-task and off-task thoughts but that there would be many instances in which the subject suffered a performance lapse despite a self-report of being "on task." Unfortunately, with current methods we cannot objectively assess the accuracy of participants' reports of subjective states (e.g., "I am mind-wandering."). Thus, in the second part of our study (Experiments 3a and 3b), we instead probed subjects' awareness of performance fluctuations (e.g., number of items held in mind on each trial). This approach allowed us to compare the number of confident responses with the number of correct responses on a given trial, with the prediction that confidence may still be high on performance lapse trials. Further, by probing metaknowledge on each trial, we could test whether high- and low-capacity individuals also differ in their metaknowledge accuracy.

## Experiment 1

In Experiment 1, we probed participants about the content of their thoughts during a whole-report working memory task (Adam et al., 2015; Huang, 2010). In the whole-report task, subjects are shown an array of colored squares and then, after a brief blank delay period, are asked to report the colors of each of the items from the array by clicking on a color patch at the location of each item in any order that they choose. A critical advantage of this working memory task is that it provides graded information about the subject's performance on each trial, allowing us to measure performance fluctuations throughout the session. We defined performance lapses as trials in which the number correct was well below typical estimates of capacity (i.e., zero or one correct). Consistent with this definition of performance lapses, formal models of change-detection performance have used error rates for Set Size 2 arrays to determine an attention lapse rate (Rouder et al., 2008), and previous modeling work for the whole-report task found that zero or one correct is indistinguishable from random guessing for large (six-item) arrays (Adam et al., 2015).

After a random subset of whole-report trials, we asked participants to categorize the contents of their thoughts as either "on-task" or as one of three types of off-task thoughts (mind-wandering, task-related interference, external distraction). We had two main aims for this experiment. First, we wanted to replicate the finding that subjective ratings of thought content predict trial-by-trial visual working memory performance (Unsworth & Robison, 2016). Second, we used a strong difficulty manipulation to test the role of task difficulty on mind-wandering rates. In Experiment 1, trials were subdivided into blocks of "easy" and "difficult" memory array sizes. During easy blocks, participants were asked to remember arrays that were within typical working memory capacity limits (two to three items). During difficult blocks, participants were consistently asked to remember arrays that far exceed typical working memory capacity limits (six to eight items).

### Materials and method

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study (Simmons, Nelson, & Simonsohn, 2012). In addition, all raw data are available on Open Science Framework (https://osf.io/syv5w/).

**Participants** There were 41 participants in Experiment 1. For this and all other experiments, we aimed for a sample size of between 30 and 50 participants, and stopped data collection at convenient time points (e.g., at the end of the academic term) before analyzing the data. All participants gave written informed consent according to procedures approved by the University of Oregon institutional review board. All participants had normal color vision and normal or corrected-to-normal visual acuity. Participants were compensated for participation with course credit or monetary payment ($8/hour).

**Stimuli** Stimuli were generated in MATLAB (The MathWorks, Natick, MA) using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). Participants sat approximately 60 cm from a 17-inch flat CRT monitor (refresh rate of 60 Hz) in a dimly lit room. Colored squares (~2.5° visual angle) were presented on a gray background (RGB = 128 128 128). Nine distinct colors were used for all experiments (RGB values: Red = 255 0 0; Green = 0 255 0; Blue = 0 0 255; Magenta = 255 0 255; Yellow = 255 255 0; Cyan = 0 255 255; Orange = 255 128 0; White = 255 255 255; Black = 0 0 0). Participants were instructed to fixate a small white dot (~.25° visual angle) at the center of the display.

**Procedures** Participants completed a whole-report memory task. On each trial of the whole-report task, participants briefly

viewed (250 ms) an array of colored squares. After a blank delay (1,000 ms), participants recalled each item from the memory array. At response, participants were shown a 3 × 3 grid of colors at the location of each memory array item. Participants were instructed to use a mouse to click the color in the grid corresponding to the remembered item at each location. Participants could report the items in any order they wished, but they were required to respond to all items before moving on to the next trial. The next trial began after an intertrial interval of 1,000 ms. There were two task difficulty conditions. In the easy condition, all trials were Set Sizes 2 and 3; in the hard condition, all trials were Set Sizes 6 and 8. Task difficulty was blocked and interleaved such that all odd blocks were easy and all even blocks were hard.

On a randomly chosen 20% of trials, participants were probed about the content of their thoughts via a text display on the computer monitor. Performance for trials immediately preceding probes was used for analysis of the effects of thought content on working memory performance. During probes, participants were instructed to categorize their thoughts throughout the trial they had just completed into one of four categories (Stawarczyk, Majerus, Maj, Van der Linden, & D'Argembeau, 2011):

1. Totally focused on completing the task (on task)
2. Thinking about task performance (task-related interference)
3. . . . something other than the task (mind-wandering)
4. . . . something in my immediate environment (external distraction)

Before the experiment began, the experimenter gave examples, explained in detail, and checked for participant understanding of these four categories. To respond, participants pressed the number on the keyboard that best corresponded to their thought contents.

Participants completed 10 blocks in total of the whole-report memory task. A total of 21 participants completed 10 blocks of 20 trials (200 trials total, 40 probed trials), and 20 participants completed 10 blocks of 30 trials (300 trials total, 60 probed trials). To preview results, there was no difference in performance for these two groups of participants ($p > .7$), and all reported results are combined across all participants.[1]

[1] An uneven number of trials could potentially introduce a problematic confound. For example, if an effect is driven only by the trials at the very end of an experiment (i.e., the last one third of trials), then these effects might be driven only by the participants with more trials. If there is a confound, then the number of trials should be made equivalent across conditions. On the flipside, if increasing the number of trials simply increases the signal-to-noise ratio (and the reliability of the observed effects) and does not introduce a confound, then it would be beneficial to keep all trials for all subjects. As a check, we reran all analyses with an equal number of trials for all subjects. We found no differences in any of the patterns of results, so we decided to keep all trials to maximize reliability.

## Results

On average (across both probed and unprobed trials), participants correctly reported 2.29 items in the easy blocks ($SD$ = .14, ceiling = 2.5 correct) and 2.91 items in the hard blocks ($SD$ = .49, ceiling = seven correct). This difference was significant, $t(40) = 10.05$, $p < .001$, 95% CI [.5, .75], and likely due to a ceiling of 2.5 items in the easy condition. We also examined lapse rate, defined as the proportion of trials in which participants correctly reported zero or one items. Participants experienced fewer lapses during easy blocks ($M$ = 7.1%, $SD$ = 5.4%) than hard blocks ($M$ = 11.4%, $SD$ = 8%), and this difference was significant, $t(40) = 5.6$, $p < .001$, 95% CI [3%, 6%]. There was no significant difference in mean performance for probed versus unprobed trials ($p = .97$). Likewise, there was no difference in lapse rate for probed versus unprobed trials ($p = .34$).

Participants reported that they were on task 44% of the time ($SD$ = 25%), experiencing task-related interference 24% of the time ($SD$ = 15%), mind-wandering 27% of the time ($SD$ = 20%), and experiencing external distraction 5% of the time ($SD$ = 6%; see Fig. 1a). Because we grouped set sizes into easy and hard blocks, we checked to see whether probe responses changed as a function of task difficulty (see Fig. 1b). Participants reported significantly fewer on-task thoughts in hard blocks ($M$ = 28%, $SD$ = 28%) than in easy blocks ($M$ = 61%, $SD$ = 25%), $t(40) = 10.34$, $p < .001$, 95% CI [26%, 39%]. Likewise, rates of all three "off-task" categories increased. Participants reported significantly more task-related interference in hard blocks ($M$ = 30%, $SD$ = 21%) than in easy blocks ($M$ = 17%, $SD$ = 16%), $t(40) = -3.75$, $p < .001$, 95% CI [−20%, −6%]. Likewise, there was more mind-wandering in hard blocks ($M$ = 35%, $SD$ = 25%) than in easy blocks ($M$ = 19%, $SD$ = 19%), $t(40) = -4.97$, $p < .001$, 95% CI [−23%, −10%]. We did not replicate the finding that mind-wandering rates predict working memory performance. Across all trials, the correlation between mind-wandering and mean performance was $r = -.07$, $p = .67$. The correlation was numerically stronger in the predicted direction for hard trials ($r = -.14$, $p$ = .37) relative to easy trials ($r = .08$, $p = .61$), but still not significant.[2] However, given the typical correlation strength of around $r = -.3$ that is found in the literature, we would have needed 70 subjects to detect this effect with 80% power. Finally, participants more frequently reported thinking about external distractions in hard blocks ($M$ = 7%, $SD$ = 8%) than in easy blocks ($M$ = 3.5%, $SD$ = 6%), $t(40) = -2.6$, $p = .0123$, 95% CI [−6%, −1%].[3]

Next, we examined whether memory performance changed as a function of thought content. Figure 2a shows

[2] Figures for these and all other reported between-subjects correlations are available in our Supplementary Materials section.

[3] Corrected $p$ value for $\alpha = .05$ with four multiple comparisons is $p = .0125$.

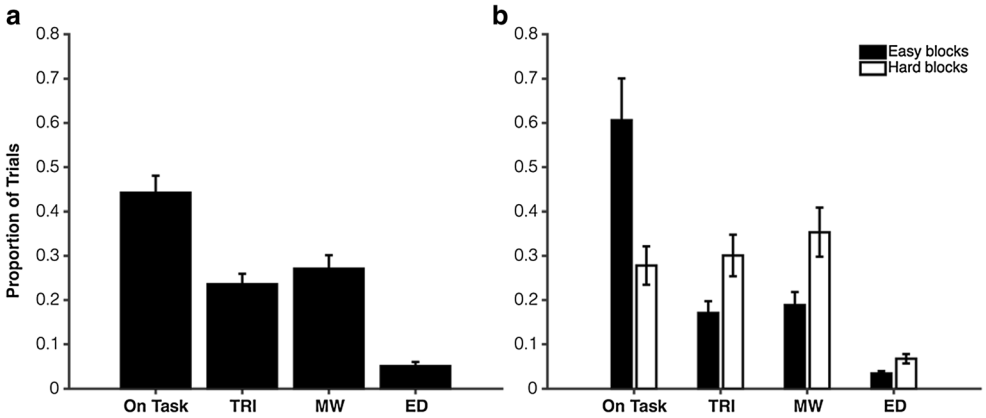


**Fig. 1** Distribution of thought-probe responses in Experiment 1. *TRI* = task-related interference; *MW* = mind-wandering; *ED* = external distraction. **a** Proportion of responses for all trials. **b** Proportion of responses separated by easy blocks (Set Sizes 2 and 3) and hard blocks (Set Sizes 6 and 8)

the mean number correct as a function of probe response. Unfortunately, not all participants used all four probe types within both the easy and difficult conditions. For example, only 17 out of 41 participants reported external distraction during the easy condition. Because of unequal numbers of trials, we initially conducted a series of pairwise comparisons to examine performance as a function of thought category within each difficulty condition. For each pairwise comparison, we included only participants who made responses in the two categories being compared (range: 15–40 participants). Results of the comparisons are shown in Tables 1 and 2. Overall, there was no significant modulation of mean performance in the easy blocks (all *p*s > .30). In the hard blocks, performance was higher in the on-task category compared to all three off-task categories (all *p*s < .02). However, there was no significant performance difference between the three off-task categories (all *p*s > .25).

To more rigorously assess within-subject changes, we collapsed task-related interference, mind-wandering, and external distraction into the category "off task" and ran a two-way repeated-measures ANVOA with Difficulty (easy versus difficult) and Task State (on task vs. off task) as factors. There were 27 participants total who had responses in all four categories: (1) easy blocks, on task; (2) easy blocks, off task; (3) difficult blocks, on task; and (4) difficult blocks, off task. Only these 27 participants were used in the analysis. The other 14 participants had zero trials in any of the four categories.

We first examined average performance (mean number of items correctly identified). Results are depicted in Fig. 2b. There was a significant main effect of both Difficulty, $F(1, 26) = 41.4$, $p < .001$, $\eta_p^2 = .61$, and Task State, $F(1, 26) = 11.5$, $p = .002$, $\eta_p^2 = .31$, as well as a significant Difficulty × Task State interaction, $F(1, 26) = 10.8$, $p = .003$, $\eta_p^2 = .29$. Follow-up comparisons revealed that there was no effect of Task State on mean performance in the easy blocks, $t(26) =$

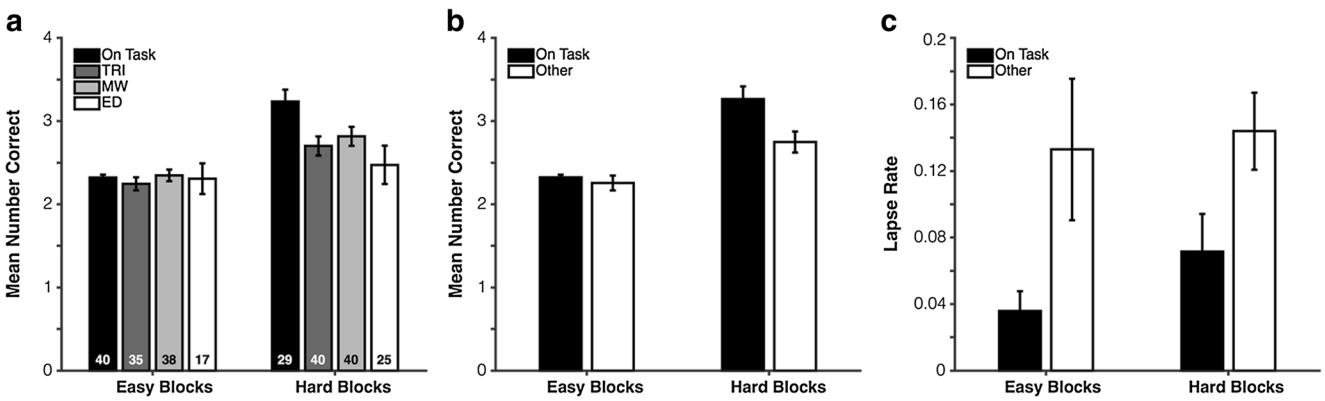


**Fig. 2** Performance as a function of thought-probe response in Experiment 1. *TRI* = task-related interference; *MW* = mind-wandering; *ED* = external distraction. **a** Mean number correct as a function of thought-probe response. Not all participants used all four response categories in the easy and hard conditions. Each bar is calculated separately using only participants who used each category. *Digits* represent the number of participants contributing to each bar, and *error bars* represent one standard error of the mean. **b** Mean number correct as a function of thought-probe response. Here, all three off-task probes are collapsed into the category *other*. Only participants who contributed to all four categories ($N = 27$) are included in the graph. **c** Lapse rate as a function of thought-probe response. Again, only participants who contributed trials to all four categories ($N = 27$) are included in the graph

**Table 1** Experiment 1, easy condition: Pairwise comparisons for mean accuracy as a function of thought-probe type

| | Task-related interference | Mind-wandering | External distraction |
|---|---|---|---|
| *On task* | $p = .38$ | $p = .83$ | $p = .86$ |
| *Task-related interference* | – | $p = .64$ | $p = .91$ |
| *Mind-wandering* | – | – | $p = .91$ |

.69, $p = .50$, but there was a strong effect of Task State in the hard condition, $t(26) = 4.4$, $p < .001$.

Finally, we looked specifically at the rate of lapses (trials where subjects got zero or one items correct) as a function of probe response (see Fig. 2c). There was no significant main effect of Difficulty on lapse rate, $F(1, 26) = 1.05$, $p = .32$, $\eta_p^2 = .04$. There was, however, a main effect of Task State, $F(1, 26) = 8.5$, $p = .007$, $\eta_p^2 = .25$. There was no significant Difficulty × Task State interaction, $F(1, 26) = .25$, $p = .62$, $\eta_p^2 = .01$. Despite the significant difference in lapse rate when considering all trials, there was no main effect of difficulty on lapse rate after dividing trials by task state. This suggests that although lapse rate differed significantly across difficulty conditions, participants caught their lapses at about the same rate, regardless of the relative preponderance of lapses within each condition.

## Discussion

As predicted, working memory performance varied with subjective reports of task-unrelated thoughts, but many working memory failures persisted even when subjects reported being on task. Reports of off-task thoughts were associated with lower average working memory performance, but only in the difficult condition. There was a strong ceiling effect in the easy task condition, so we may have seen no difference in average performance in the easy condition largely because of restricted range. Indeed, when we instead assessed the rate of extreme failures (zero or one correct), for which there was no ceiling effect, we found a significant relationship between performance and thought content for both easy and difficult conditions. Interestingly, when participants reported that they were "on task," they still experienced a large number of working memory failures. Again, working memory failures were defined as trials for which participants performed well below capacity limits (zero or one items correct). Failure trials were indeed far less frequent when participants reported being on task (~5%) relative to off task (~14%), it is nevertheless striking that many failure trials (i.e., a third of the typical rate) persisted when participants reported they were completely focused on the task.

Mind-wandering rates increased dramatically during the difficult condition; the proportion of on-task thoughts decreased by around half. This striking result is inconsistent with previous work showing no relationship between working memory load and mind-wandering rates (Mrazek et al., 2012; Teasdale, Proctor, Lloyd, & Baddeley, 1993; Unsworth & Robison, 2016) or decreased mind-wandering for higher working memory loads (Teasdale et al., 1993). This effect is also inconsistent with mind-wandering rates during attention tasks (e.g., SART), which almost universally show decreased mind-wandering with increased task difficulty (e.g. Antrobus, 1968; McKiernan, D'Angelo, Kaufman, & Binder, 2006). Despite this consistent trend, there is reason to believe that complex tasks may affect mind-wandering differently than simple sustained attention tasks that are typically used in the mind-wandering literature; in one study of real-world mind-wandering, low-capacity individuals in particular reported more frequent mind-wandering when engaged in challenging real-world tasks (Kane et al., 2007). In addition, we think that increased mind-wandering for very difficult tasks could make sense in light of an executive-failure view of both mind-wandering and working memory performance. In our study, the difficult condition placed a heavy burden on executive resources. The memory load was well above working memory capacity, and trials were consistently difficult. In contrast, Teasdale and colleagues (1993) ensured that all participants performed near perfect on the memory tasks. Unsworth and Robison (2016) included some supracapacity set sizes, but the trials were relatively fast-paced, and difficulty levels were intermixed. There were some difficult set sizes, but these were relatively infrequent; participants may have used the easy trials to "take a break" and better prepare for upcoming difficult trials.

**Table 2** Experiment 1, difficult condition: Pairwise comparisons for mean accuracy as a function of thought-probe type

| | Task-related interference | Mind-wandering | External distraction |
|---|---|---|---|
| *On task* | **$p = < .001$** | **$p = .004$** | **$p = .017$** |
| *Task-related interference* | – | $p = .37$ | **$p = .29$** |
| *Mind-wandering* | – | – | **$p = .56$** |

Significant comparisons are in bold typeface

Difficult working memory tasks pose several challenges for the validity of thought probe ratings. One important potential alternative explanation of the relationship between working memory performance and thought content is that participants reported perceived performance rather than the content of their thoughts (Head & Helton, 2016). That is, during more difficult trials, participants may have reported experiencing "off-task thoughts" as an excuse for performing poorly. Because the trials were blocked by difficulty, this task performance bias may have been particularly pronounced. In addition, the blocking of difficulty conditions could have affected subjects' ratings in other ways (e.g., participants may dislike hard blocks). While the present data would be consistent with increased mind-wandering during difficult working memory blocks, future experiments are needed to provide further support for this relationship. In particular, it will be important to look at mind-wandering rates across a wide variety of tasks (both easy and difficult) to decouple mind-wandering rates from trial-specific or task-specific performance.

While it is important to question the validity of subjective ratings made immediately after a recall screen, we think it is possible that the subjective judgments made in Experiment 1 reflect more than just perceived performance. First, the difference in average performance for on-task versus off-task thoughts was relatively small in magnitude (~2.8 vs. 2.5 items, respectively), indicating that participants were still performing well even when they reported being off task. That is, if participants were attempting to use thought probes to indicate their level of performance, they were not doing a very accurate job of it. Second, the rate of performance failures has been shown to increase substantially as a function of memory load (Adam et al., 2015), which may result in part from a concomitant increase in off-task thoughts for supracapacity trials. Finally, participants were given no feedback about performance. Even if we assume the pessimistic position that thought content judgments solely reflect participants' perceived performance, we could still conclude that (1) subjective judgments predict trial-by-trial lapses of working memory performance and (2) despite this reliable introspection, many working memory failures go undetected.

## Experiment 2

In Experiment 1, participants binned their thoughts into one of four discrete categories. However, some inattentive states might not be well described by one of these four thought categories (e.g., zoning out). In Experiment 2, we instead had participants rate their subjective attention state on a continuous scale from 1 to 9, with 1 being the most off task and 9 being the most on task (Unsworth & McMillan, 2014a, 2014b). We predicted that both mean working memory performance and working memory failure rates would covary with subjective ratings of attention state. In this experiment, we also included easy trials (two items) and difficult trials (six items). Instead of blocking difficulty conditions, trial difficulty was varied randomly from trial to trial in order to test whether or not the blocking of difficulty accounted for the large increase in off-task thoughts for difficult trials in Experiment 1.

### Materials and method

**Participants** There were 34 participants in Experiment 2. All participants gave written informed consent according to procedures approved by the University of Oregon institutional review board. Participants were compensated for participation with course credit or monetary payment ($8/hour). All participants had normal color vision and normal or corrected-to-normal visual acuity. Two participants were excluded for reporting the same attention state on every trial, and two participants were excluded for task noncompliance. This left a total of 30 participants for analysis.

**Stimuli** Stimuli were identical to Experiment 1, with one exception. Upon responding to an item, the 3 × 3 response matrix for that item would desaturate. Desaturated RGB values were as follows: Red = 255 153 153; Green = 153 255 153; Blue = 153 153 255; Magenta = 255 153 255; Yellow = 255 255 153; Cyan = 153 255 255; Orange = 255 204 153; White = 255 255 255; Black = 110 110 110.

**Procedures** Participants completed nine blocks of 32 trials of the whole-report memory task (288 trials total). Memory arrays were either Set Size 2 or Set Size 6. Set Size 2 trials will be referred to as "easy" trials, and Set Size 6 trials will be referred to as "hard" trials. Trial difficulty was randomized from trial to trial. Items were presented for 200 ms and remembered across a blank delay of 1,000 ms. Participants could report the items in any order they wished, but they were required to respond to all items before moving onto the next trial. The next trial began after an intertrial interval of 1,000 ms.

After a randomized 25% of trials, participants were probed about their current level of attention (72 probed trials total). They were asked to rate their attention on a scale from 1 to 9, with 1 meaning *not at all focused on the current task* and 9 meaning *completely focused on the current task*. Before the experiment began, the experimenter explained the ratings and checked for participant understanding. To respond to the probe, participants pressed the number on the keyboard that best corresponded to their current attention state at the moment of the probe. Trials immediately preceding probes were used for analysis of the relationship between working memory performance and attention state.

## Results

On average (across both probed and unprobed trials), participants correctly reported 1.84 items on easy trials ($SD$ = .11, ceiling = two correct) and 2.66 items on hard trials ($SD$ = .44, ceiling = six correct). This difference was significant, $t(29)$ = 10.74, $p < .001$, 95% CI [.66, .97], and likely due to a ceiling effect in the easy condition (maximum = 2.0). Participants experienced fewer lapses on easy trials ($M$ = 13.3%, $SD$ = 7.5%) compared to hard trials ($M$ = 16.3%, $SD$ = 8.6%), and this difference was significant, $t(29)$ = 2.15, $p$ = .04, 95% CI [.1%, 6%]. There was no significant difference in mean performance for probed versus unprobed trials ($p$ = .40) and no difference in lapse rate for probed versus unprobed trials ($p$ = .90).

Participants reported being slightly more on-task on easy trials ($M$ = 5.73, $SD$ = 1.75) compared to hard trials ($M$ = 5.18, $SD$ = 1.71), $t(29)$ = 3.26, $p$ = .003, 95% CI [.21, .90]. Distributions of attention state ratings are shown in Fig. 3a; average attention state is shown in Fig. 3b. Not all participants used the entire range of the attention state scale. Because of this, we used a linear mixed-effects model with Subject entered as a random factor. A linear mixed-model approach has been standard for attention state ratings of this kind (Unsworth & McMillan, 2014a, 2014b) because of robustness to unbalanced designs and missing data (Kliegl, Wei, Dambacher, Yan, & Zhou, 2011). First, we examined the relationship between attention state and mean number correct. Because of the strong ceiling effect for easy trials, we ran separate models for easy trials and difficult trials. Each model included Attention State as a fixed factor and Subject as a random factor. The model for easy trials revealed a significant positive relationship between mean number correct and attention state, $t$ = 4.47, $p < .001$ ($b$ = .03, $SE$ = .006). Likewise, there was a significant positive relationship between mean number correct and attention state for difficult trials, $t$ = 6.82, $p < .001$ ($b$ = .14, $SE$ = .02). Mean number correct as a function of attention state is illustrated in Fig. 4a.[4]

Next, we examined the relationship between lapse frequency and attention state. We first ran linear mixed-effects model with Attention State and Difficulty as fixed factors and Subject as a random factor. There was a negative relationship between attention state and lapse rate, $t$ = −2.50, $p$ = .012 ($b$ = −.018, $SE$ = .007). There was only a marginal effect of task difficulty, $t$ = 1.75, $p$ = .081 ($b$ = .017, $SE$ = .010) on lapse rate, and no significant Difficulty × Attention State interaction, $t$ = −1.24, $p$ = .22 ($b$ = −.002, $SE$ = .002). As such, we ran a second model collapsing across the two difficulty levels. The increased number of trials led to a stronger estimate of attentional state on lapse rate, $t$ = −6.79, $p < .001$ ($b$ = −.03, $SE$ = .004). Lapse rate as a function of attention state is illustrated in Fig. 4b.

## Discussion

Previously, subjective measures of attention state have been shown to correlate with trial-by-trial performance in measures of goal-neglect and fluid intelligence (Unsworth & McMillan, 2014a, 2014b). Here, we found that subjective ratings of attention state predicted trial-by-trial working memory performance. When participants rated their attention state as high (more on task), they had higher average working memory performance and were far less likely to have a lapse in performance. Nevertheless, as in Experiment 1, there was still a nearly 10% lapse rate observed even when participants reported being near the top of the attention rating scale. The average lapse rate across all trials was 15%, meaning that participants in this experiment only noticed an average of approximately one third of their lapses. Thus, despite some accurate metaknowledge about overall performance, performance failures went undetected more often than they were caught.

We found a small difference in task-unrelated thoughts as a function of memory load, even though set sizes were intermixed. After hard trials, participants on average rated their attention state as slightly lower than after easy trials. From these data alone it is not possible to say whether average attention state was lower after hard trials because participants truly experienced more lapses of attention during hard trials or because participants reported perceived performance. Indeed, it is a bit puzzling that we found an effect of set size on attention ratings even though trial difficulties were intermixed in this experiment. If fluctuations of attention are randomly interspersed across the session, then high and low attention state ratings should be distributed equally among easy and difficult set sizes, and there is evidence in the literature to support this intuition. The closest data set to our own is from Unsworth and Robison (2015), in which they had participants report mind-wandering during a change-detection task. Unlike the present study, Unsworth and Robison found no relationship between trial difficulty (Set Sizes 1–8) and mind-wandering rates. Assuming that people have equally good metaknowledge in both tasks (e.g., whether or not colored square $X$ was in memory), it is then surprising that we found differences in mind-wandering as a function of set size when Unsworth and Robison did not. That is, accuracy is much lower on high set-size change detection trials. So if participants report their perceived accuracy in their thought probe responses, then Unsworth and Robison should also have observed the relationship between thought content and set size. Given that they did not observe this, we speculate that the whole-report response may explain the small difference in

[4] By eye, it looks like the "1" ratings may explain the positive relationship, particularly for the easy condition. However, after excluding all attention state ratings of 1 from the mixed-effects model, there was still a significant positive relationship between attention state and mean number correct in both the easy condition, $t(407.9)$ = 2.97, $p$ = .003 ($b$ = .03, $SE$ = .007), as well as in the difficult condition, $t(563.4)$ = 4.57, $p < .001$, ($b$ = .11, $SE$ = .02).

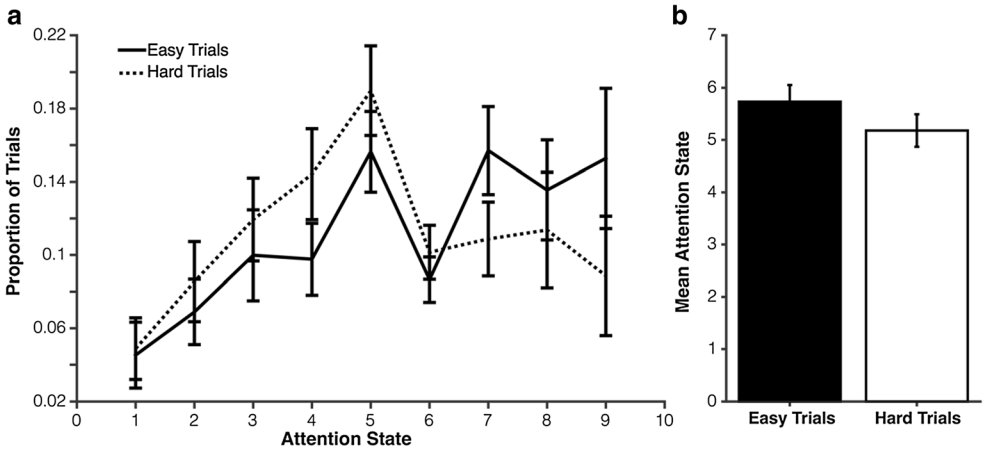


**Fig. 3** Distribution of attention state ratings by condition in Experiment 2. *Error bars* represent one standard error of the mean. Attention state ratings reflected the degree to which participants felt they were focused on completing the task at hand, with 1 meaning *not at all focused on the task* and 9 meaning *totally focused on completing the task.* **b** Distribution of attention state ratings as a function of set size (easy vs. hard). **b** Average attention state rating as a function of trial difficulty

attention state ratings between set sizes. Specifically, because of the whole-report nature of the task, participants took longer to respond to difficult trials than to easy trials (~7 seconds vs. ~2 seconds). As such, it is possible that participants were more likely to become inattentive during this longer response period because of prolonged cognitive demands.

## Experiment 3

The thought probes used in Experiments 1 and 2 have two major shortcomings. First, each experiment had only a small number of probed trials. To be consistent with the existing literature on task-unrelated thoughts, we chose only to probe participants about their thoughts on a small subset of trials. However, because we only probed a small percentage of trials, we could not take full advantage of the trial-by-trial resolution afforded by the whole-report working memory measure. Second, we could not objectively measure the accuracy of subjects' meta-awareness of internal states. Instead, we had to take participants' ratings of their internal states at face value.

In Experiments 3a and 3b, we instead had observers report subjective confidence for each item that they reported. By

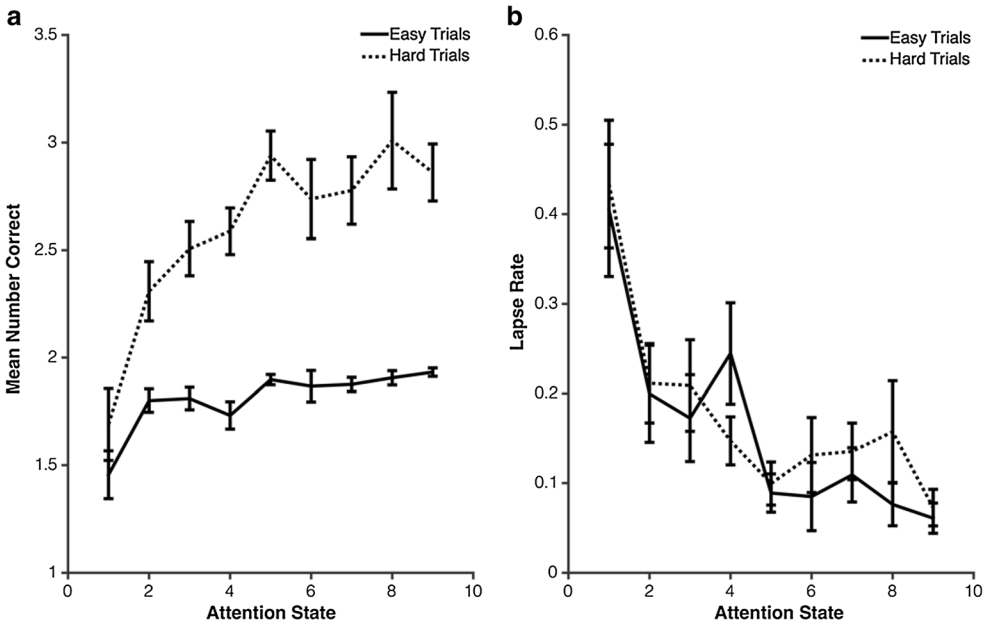


**Fig. 4** Performance as a function of attention state rating in Experiment 2. *Error bars* represent one standard error of the mean. *Attention state* ratings reflected the degree to which participants felt they were focused on completing the task at hand, with 1 meaning *not at all focused on the task* and 9 meaning *totally focused on completing the task. Solid lines*: Easy trials (Set Size 2). *Dotted lines*: Hard trials (Set Size 6). **a** Mean number correct as a function of attention state rating. **b** Lapse rate (0 or 1 correct) as a function of attention state rating

collecting both confidence ratings and accuracy for every item and every trial, we had more power to examine trial-by-trial relationships between accuracy and confidence. Further, because subjective ratings were on the same scale as accuracy (number of items), we could directly measure bias in metacognition. Because participants had some number of working memory failures even when they reported being fully attentive, we predicted that participants would have a positive bias in confidence ratings, particularly for failure trials. We further predicted that individuals with poor working memory performance would suffer the "dual burden" of poor metacognitive insight (Kruger & Dunning, 1999).

In Experiment 3a, we repeated the same challenging set size (six items) for a large number of trials (300). We collected both accuracy and confidence ratings for each item in order to examine trial-by-trial fluctuations in working memory performance. Once again, participants could report the items in any order they chose. In Experiment 3b, we replicated the manipulation in Experiment 3a and also added a control condition in which the computer randomly determined the order in which participants must report the items. This random-response order condition allowed us to estimate and control for the effects of output interference in Experiment 3a.

## Materials and method

**Participants** There were 45 participants in Experiment 3a and 38 in Experiment 3b. One subject was excluded from Experiment 3a for failure to comply with task instructions, leaving 44 participants for analysis. Four participants were excluded from Experiment 3b for the following reasons: failing to complete both tasks (one subject), chance-level performance (one subject), or failure to comply with task instructions (two subjects). Some aspects of the data from Experiment 3a have been previously reported (Adam et al., 2015, Experiment 1b), but all analyses presented here are novel. Participants in both experiments also completed a color change detection task at the end of the experiment (results not reported in this study).

**Stimuli** Stimuli and timing parameters were identical to those in Experiment 1. In the random response-order condition of Experiment 3b, the to-be-reported square was indicated by a light gray box drawn around the response pad (RGB = 170 170 170).

**Procedures for Experiment 3a** Participants completed 10 blocks of 30 trials (300 trials total); all arrays were Set Size 6, and colors were chosen without replacement from the set of nine possible colors. By using arrays that were only one set size, we could examine fluctuations in performance that were disentangled from differences in difficulty from trial to trial. At test, participants could report the items in any order they chose. While responding, participants were instructed to report their confidence in each response by using the left and right mouse buttons. Participants were instructed to click their color choice with the left mouse button if they felt they had any information in mind about the color of the item. Likewise, they were instructed to click their color choice with the right mouse button if they felt they had no information in mind about the color of the item.

**Procedures for Experiment 3b** Participants completed two conditions of the whole-report task (60 trials per condition): free response order and random response order. The order of the two conditions was counterbalanced across participants. As in Experiment 3a, all arrays were Set Size 6, and colors were chosen without replacement from the set of nine possible colors. The free response-order condition was identical to Experiment 1a; participants were allowed to report the six items in any order they wished. In the random response-order condition, participants instead had to report the items in an order dictated by the computer. At the beginning of the response period, the computer indicated which item must be reported by drawing a light gray frame around the item. After the participant responded to the probed item, the computer moved the frame to the next to-be-reported item. This process was repeated until the subject had made a response for every item. In both conditions, participants reported confidence in each item using the left and right mouse buttons as in Experiment 3a.

## Results

**Experiment 3a** On average, participants correctly identified an average of 2.88 items ($SD$ = .49), and they reported being confident about 3.04 items ($SD$ = .52) out of six possible items. There was no significant difference between the mean number of correct items and the mean number of confident items, $t(43) = 1.64$, $p = .11$, 95% CI [−.04, .36]. However, looking at the full distribution of responses reveals some systematic differences in the underlying distribution of confident responses relative to correct responses (see Fig. 5a). Specifically, participants seem to have overreported their modal performance outcome (three items).

In addition to looking at total trial performance, we can look at confidence and accuracy for each individual response within the trial. All trials were Set Size 6, so participants made six responses total. Figure 5b shows proportion correct and confident as a function of response number for all trials. As participants were free to report the items in any order they chose, performance and confidence were initially high (for the first three responses) and then dropped precipitously at Response 4. On lapse trials (zero or one correct), however, there was a stronger disconnect between performance and confidence.

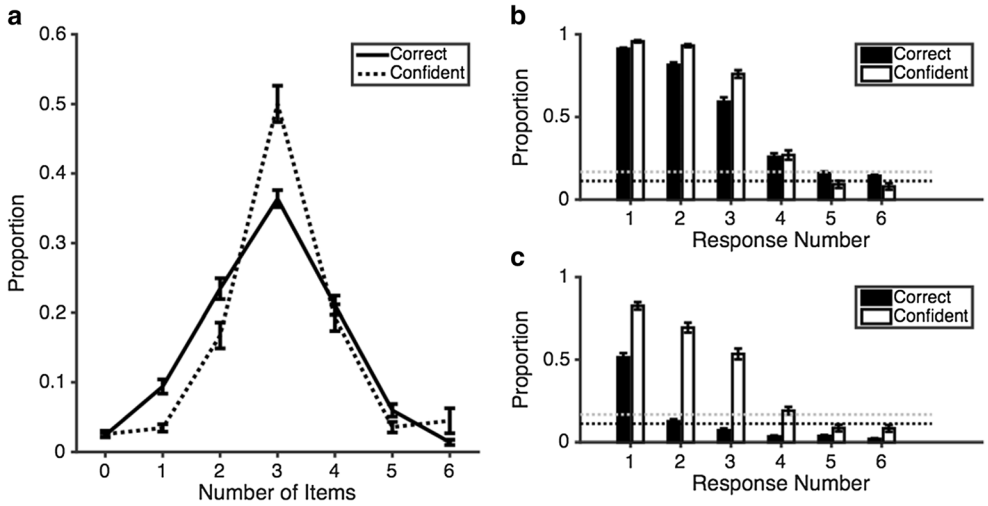


**Fig. 5** The relationship between correct and confident responses in Experiment 3a. *Error bars* represent one standard error of the mean. **a** *Dotted line*: Proportion of trials where participants reported that they were confident about zero through six items. *Solid line:* Proportion of trials where subject correctly reported zero through six items. **b** Distribution of correct and confident responses across each response in time across all trials. *Response number* = 1 represents the first item the subject reported. *Response number* = 6 represents the last item the subject reported. The *gray dotted line* represents a "smart" guessing strategy of remembering the colors of three items and guessing only among the six possible nonremembered colors (1/6), and the *black dotted line* represents a "purely random" guessing strategy among all possible colors (1/9). **c** Distribution of correct and confident responses across each response in time only for lapse trials (participants got a total of zero or one items correct)

Here, accuracy was above chance for the first response but quickly fell to below-chance levels for later responses. Despite this pattern of performance, participants still reported that they were confident in the first three responses.

Next, we wanted to more formally test the predictions that (1) there is a reliable trial-by-trial relationship between accuracy and confidence and (2) despite this reliable relationship, participants underestimate failures (zero or one correct). For each individual subject, we calculated the correlation coefficient between number of correct responses and number of confident responses. The average correlation value was $r$ = .34 ($SD$ = .16, average $p < .05$), and 40 out of 44 participants had statistically reliable within-subject correlations ($p < .05$). To quantify awareness of failure trials, we calculated a lapse sensitivity measure (lapses detected / total number of lapses). That is, of all the trials in which participants got zero or one items correct, what proportion of the time did they report that they were confident on zero or one items? Average sensitivity was only .28 ($SD$ = .19), indicating that participants accurately caught extreme failures only about a quarter of the time. While $d$-prime is a more commonly used means of quantifying discriminability, we could not use this metric because of a number of participants with hit rates or false alarm rates of zero (thus yielding $d$-prime values of +/− infinity). Average hit rate in Experiment 3a was 27.5% ($SD$ = 19.0%), and average false alarm rate was 3.4% ($SD$ = 4.3%).

Next, we asked whether there were systematic differences in the accuracy of metacognition as a function of overall performance. To do so, we divided participants into quartiles and examined actual performance (correct items) versus perceived performance (confident items). We ran two mixed ANOVA models using Metaknowledge (actual vs. perceived) as a within-subjects factor and Quartile as a between-subjects factor to predict (1) mean number correct and (2) lapse rate.

Consistent with the Dunning-Kruger effect, poor performers showed a larger discrepancy between perceived and actual performance (see Fig. 6). There was a significant main effect of Quartile on lapse rate, $F(3, 40) = 42.6$, $p < .001$, $\eta_p^2 = .76$. There was a significant main effect of Metaknowledge, indicating that reported lapse rates were significantly lower than actual lapse rates, $F(1, 40) = 34.2$, $p < .001$, $\eta_p^2 = .46$. Critically, there was an interaction between Metaknowledge and Quartile, indicating that the difference between perceived performance and true performance was larger for poor performers relative to good performers, $F(3,40) = 8.13$, $p < .001$, $\eta_p^2 = .38$.

We found the same effects for mean performance as for lapse rate. There was a significant main effect of Quartile on mean performance, $F(3, 40) = 15.0$, $p < .001$, $\eta_p^2 = .53$. There was a significant main effect of Metaknowledge, indicating that reported mean performance was significantly higher than actual mean performance, $F(1, 40) = 6.4$, $p = .016$, $\eta_p^2 = .14$ Finally, there was an interaction between Metaknowledge and Quartile, indicating that the difference between perceived performance and true performance was larger for poor performers relative to good performers, $F(3, 40) = 6.47$, $p = .001$, $\eta_p^2 = .33$.

We used a quartile split method to investigate the Dunning-Kruger effect because that is the prevailing standard in the literature. To supplement and strengthen this

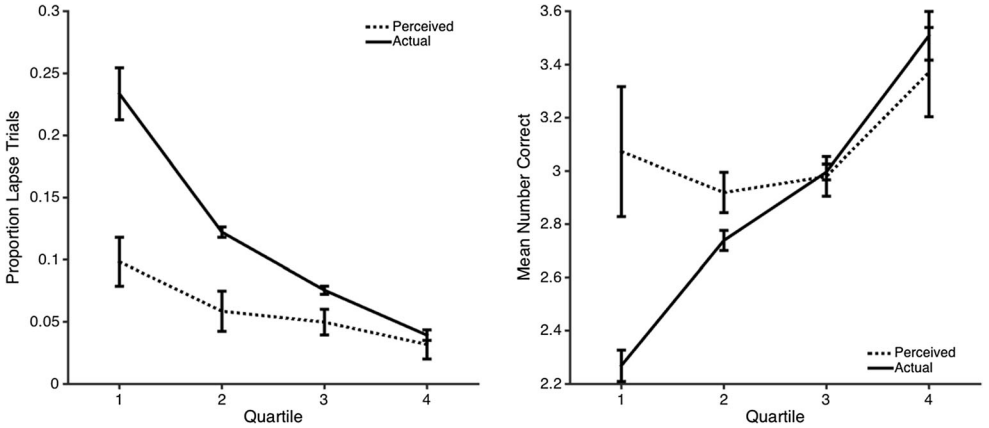


**Fig. 6** Metacognitive bias as a function of task performance in Experiment 3a. *Left:* Lapse rate (*perceived* and *actual*) as a function of task performance (quartile split). *Right:* Mean number of items correct (*perceived* and *actual*) as a function of task performance (quartile split). *Error bars* represent standard error of the mean

analysis, we computed the correlation coefficient between average performance (mean number correct) and the various metaknowledge metrics summarized above. There was a significant negative correlation between lapse awareness (actual lapse rate – perceived rate) and overall performance, $r = -.67$, $p < .001$, 95% CI [−.81, −.47], indicating that lower performing participants were more overconfident during lapses. There was also a significant correlation with mean performance awareness (mean number correct – mean number confident), $r = .59$, $p < .001$, 95% CI [.35, .75]. We also examined our metaknowledge correlation metric (correlation strength between single-trial confidence and accuracy) and our lapse sensitivity metric (percentage of lapses caught). There was a significant relationship between the metaknowledge correlation metric and average performance, $r = .47$, $p = .001$, 95% CI [.20, .67], but no relationship between lapse sensitivity and average performance, $r = .13$, $p = .39$, 95% CI [−.17, .41].

**Experiment 3b** Participants typically reported that the first three reported items were confident, and we interpreted this as evidence that participants had metaknowledge of item quality. That is, they chose to report their best remembered items first. An alternative explanation, however, could be that late responses have low accuracy only because of output interference. Therefore, participants may have reported that they were accurate early in the trial without regard to the quality of remembered items. To disentangle item-level metaknowledge from output interference, we had a new group of participants complete a free response-order condition (replicating Experiment 3a) and also complete a random response-order condition, in which the computer randomly chose the order in which participants must respond to the items.

Average performance was slightly higher during the free response-order condition ($M = 2.96$, $SD = .44$) than during the random response condition ($M = 2.58$, $SD = .61$), $t(33) = 4.98$, $p < .001$, 95% CI [.22, .53] (see Fig. 7a). The difference in

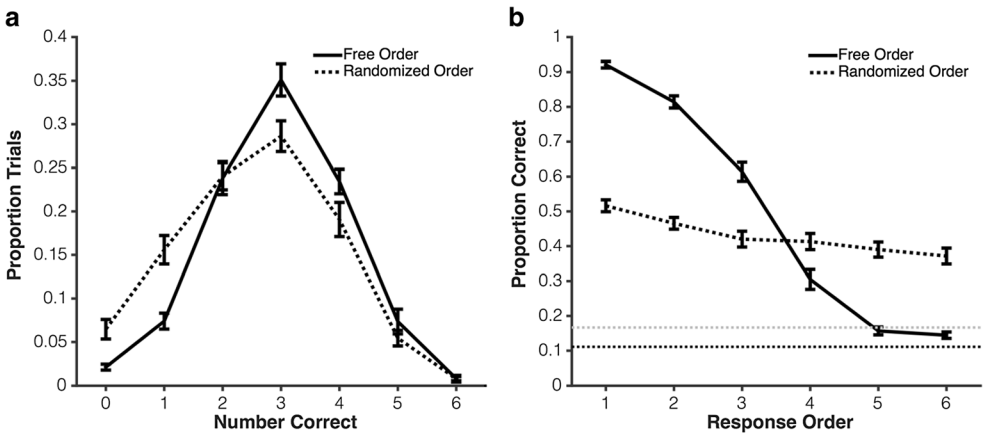


**Fig. 7** Overall performance in Experiment 3b. All *error bars* represent one standard error of the mean. The *solid line* represents performance in the free response order condition, and the *dotted line* represents performance in the randomized response order condition. **a** Distribution of performance outcomes. **b** Performance as a function of response order (1 = the first item reported, 6 = the last item reported). The *gray dotted line* represents a "smart" guessing strategy of remembering the colors of three items and guessing only among the six possible nonremembered colors (1/6), and the *black dotted line* represents a "purely random" guessing strategy among all possible colors (1/9)

accuracy for the first three responses versus the last three responses was strongly attenuated in the random response-order condition (see Fig. 7b). In the free response-order condition, participants had a mean accuracy of 78.3% (*SD* = 9.5%) on the first three responses and 20.2% (*SD* = 8.0%) on the last three responses. The average difference was 58% (*SD* = 9.6%), $t(33) = 35.4$, $p < .001$, 95% CI [55%, 61%] . On the other hand, the average difference between the first three and last three responses in the randomized order was only 7.6% (*SD* = 7.4%), $t(33) = 5.95$, $p < .001$, 95% CI [5%, 10%]. These results suggest that the decline in accuracy across responses in the free response-order condition was not due solely to output interference. Instead, this pattern of results suggests that subjects successfully stored the same number of items as in the free-recall procedure (e.g., three), but the random probing procedure distributed these accurate responses across all response positions.

Figure 8 shows performance and confidence at the trial level and at the response level in the free response-order condition. On average, participants reported that they were confident for 3.4 items (*SD* = .93) in the free-response condition, and this was significantly higher than the number of accurate items, $t(33) = 2.70$, $p = .01$, 95% CI [.11, .80]. As in Experiment 3a, participants underreported low-performance trials and overreported modal trials (three correct) and high-performance trials (six correct). When looking at responses for all trials (see Fig. 8b), confident and correct responses were both predominately early in the trial (first three responses). Likewise, on failure trials (see Fig. 8c), participants were likely to report that they were confident on the first three responses.

Figure 9 shows performance and confidence at the trial level and at the response level in the random response-order condition. On average, participants reported that they were confident about 3.1 items (*SD* = .74) in the random-response condition, and this was significantly higher than the number of correct items, $t(33) = 3.61$, $p = .001$, 95% CI [.22, .80]. At the trial level (see Fig. 9a), we once again replicated the general pattern that participants overreported modal trials and underreported poor performance trials. On the other hand, we observed that participants' confident responses were spread more evenly among response position, both for all trials (see Fig. 9b) and for lapse trials (Fig. 9c). We once again saw that participants were vastly overconfident on lapse trials (Fig. 9c), but this was not due to a response bias whereby participants always reported they were confident on the early responses. Instead, participants were confident on a specific subset of items, and the random probing procedure spread confident responses more equally across early and late responses.

We again quantified subject metaknowledge using within-subject correlations between the number confident and the number correct for each trial. In the free response-order condition, the average correlation coefficient was .29 (*SD* =.24, average $p = .16$). Twenty out of 34 participants had significant correlation coefficients. In the random response-order condition, the average correlation coefficient was .38 (*SD* = .24, average $p = .09$). Twenty-eight out of 34 participants had

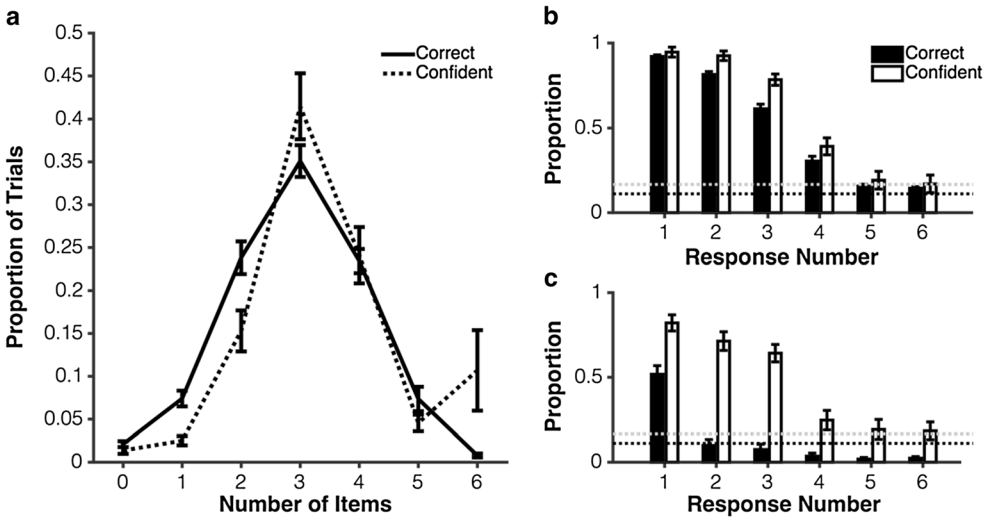


**Fig. 8** The relationship between correct and confident responses in Experiment 3b: free response order. *Error bars* represent one standard error of the mean. **a** *Dotted line*: Proportion of trials where participants reported that they were confident about zero though six items. *Solid line:* Proportion of trials where subject correctly reported zero through six items. **b** Distribution of correct and confident responses across each response in time across all trials. Response Number = 1 represents the first item the subject reported. Response Number = 6 represents the last item the subject reported. The *gray dotted line* represents a "smart" guessing strategy of remembering the colors of three items and guessing only among the six possible nonremembered colors (1/6), and the *black dotted line* represents a "purely random" guessing strategy among all possible colors (1/9). **c** Distribution of correct and confident responses across each responses in time only for lapse trials (participants got a total of zero or one items correct)

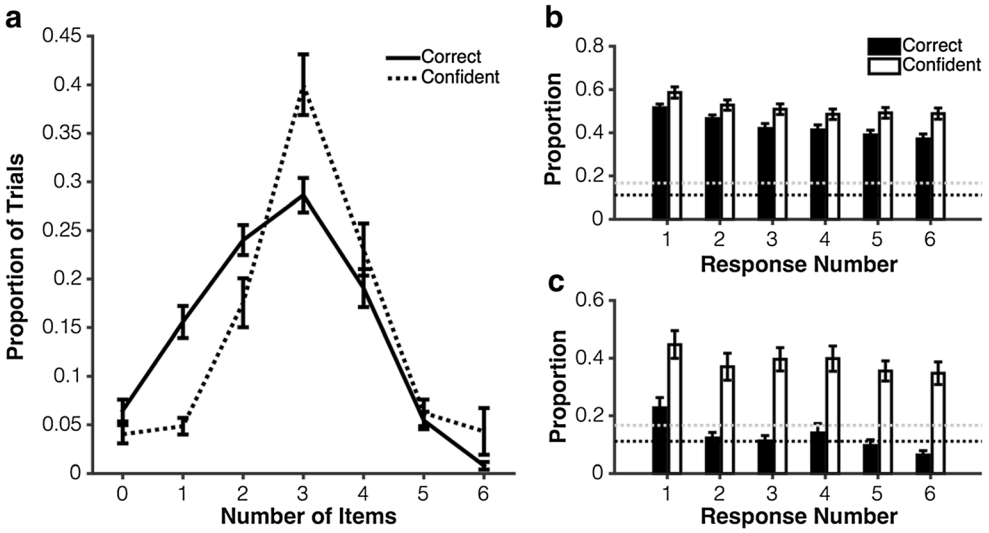


**Fig. 9** The relationship between correct and confident responses in Experiment 3b: randomized response order. *Error bars* represent one standard error of the mean. **a** *Dotted line*: Proportion of trials where participants reported that they were confident about zero through six items. *Solid line:* Proportion of trials where subject correctly reported zero through six items. **b** Distribution of correct and confident responses across each response in time across all trials. Response Number = 1 represents the first item the subject reported. Response Number = 6 represents the last item the subject reported. The *gray dotted line* represents a "smart" guessing strategy of remembering the colors of three items and guessing only among the six possible nonremembered colors (1/6), and the *black dotted line* represents a "purely random" guessing strategy among all possible colors (1/9). **c** Distribution of correct and confident responses across each responses in time only for lapse trials (participants got a total of zero or one items correct)

significant individual correlation coefficients. Note, these correlation values are numerically similar to those from Experiment 1a. However, because there were only 60 trials used to construct the correlation (as opposed to 300), relatively fewer individual participants reached the significance threshold. Combining both conditions together (120 trials total), we found an average correlation coefficient of .35 (*SD* = .23, average $p$ = .07). 29 out of 34 participants had a significant within-subject correlation between number of confident response and number of correct responses when trials from both conditions were combined. We also quantified lapse sensitivity in both conditions. In the free response-order condition, participants had an average lapse sensitivity of .22 (*SD* = .29). In the random response-order condition, participants had an average lapse sensitivity of .31 (*SD* = .29). Combined across both order conditions, lapse sensitivity was .28 (*SD* = .27). Once again, participants tended to have poor metaknowledge for extreme failure trials, noticing on average little more than a quarter.

Finally, we examined whether low performers again showed a deficit in metacognitive awareness. For this analysis, we combined trials from the free and random response-order conditions and examined metacognitive bias (perceived vs. actual performance) as a function of overall performance. We again ran mixed ANOVA models with the within-subjects factor Metaknowledge (perceived performance vs. actual performance) and the between-subjects factor Quartile to examine metacognitive bias for lapse rate and mean performance.

Despite fewer trials (120 in Experiment 3b vs. 300 in Experiment 3a), we replicated the overall pattern of results from Experiment 3a (see Fig. 10). First, we used lapse rate as our performance metric. There was a significant main effect of Quartile on lapse rate, $F(3, 30) = 27.8$, $p < .001$, $\eta_p^2 = .74$. There was also a significant main effect of Metaknowledge, indicating that perceived lapse rates were significantly lower than actual lapse rates, $F(1, 30) = 50.9$, $p < .001$, $\eta_p^2 = .63$. Critically, there was an interaction between Metaknowledge and Quartile, indicating that the difference between perceived performance and true performance was larger for poor performers relative to good performers, $F(3, 30) = 6.03$, $p = .002$, $\eta_p^2 = .38$. Second, we used mean performance as our performance metric. There was a significant main effect of Quartile on mean performance, $F(3, 30) = 7.4$, $p = .001$, $\eta_p^2 = .43$. There was a significant main effect of Metaknowledge, indicating that perceived mean performance was higher than actual performance, $F(1, 30) = 16.2$, $p < .001$, $\eta_p^2 = .35$. The interaction between Metaknowledge and Quartile was numerically similar to that observed in Experiment 3a but did not reach significance, $F(3, 30) = 1.8$, $p = .17$, $\eta_p^2 = .15$.

We again computed the correlation coefficient between average performance (mean number correct) and metaknowledge. There was once again a significant negative correlation between lapse awareness (actual lapse rate – perceived rate) and overall performance, $r = -.72$, $p < .001$, 95% CI [−.85, −.50], indicating that lower performing participants were more overconfident on lapse trials. Likewise, there was a significant correlation between overall performance awareness (mean number correct

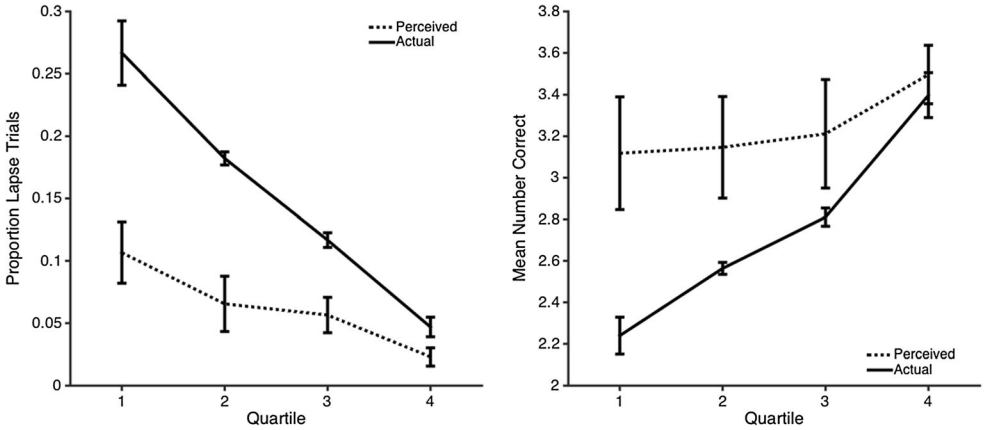


**Fig. 10** Metacognitive bias as a function of task performance in Experiment 3b. Trials were combined across the free and random conditions. *Left:* Lapse rate (*perceived* and *actual*) as a function of task performance (quartile split). *Right:* Mean number of items correct (*perceived* and *actual*) as a function of task performance (quartile split). *Error bars* represent standard error of the mean

– mean number confident), $r = .47$, $p = .005$, 95% CI [.15, .70]. We also examined our metaknowledge correlation metric (correlation strength between single-trial confidence and accuracy) and our lapse sensitivity metric (percentage of lapses caught). There was no significant relationship between the metaknowledge correlation metric and average performance, $r = .21$, $p = .23$, 95% CI [−.14, .51] or between lapse sensitivity and average performance, $r = .22$, $p = .22$, 95% CI [−.13, .52]

**Individual differences combined across Experiments 3a and 3b** We combined data across Experiments 3a and 3b in order to further illustrate individual differences in performance awareness (see Supplementary Figures S6–S9). We found a significant correlation between lapse awareness (actual lapse rate – perceived rate) and overall performance, $r = -.82$, $p < .001$, and a significant correlation between mean performance awareness (mean number correct – mean number confident), $r = -.54$, $p < .001$. In addition, we found that our correlation metric predicted overall performance, $r = .33$, $p = .003$, but our lapse sensitivity metric did not, $r = .17$, $p = .14$. To examine the robustness of these effects, we also computed the split-half reliability of each metric. We found that split-half reliability was very high for lapse awareness (perceived – actual, $r = .90$), mean performance awareness (perceived – actual, $r = .98$), and confidence-accuracy correlation strength ($r = .75$). On the other hand, split-half reliability was rather poor for the lapse sensitivity metric ($r = .48$), suggesting that it would be difficult to interpret significance of individual differences for this particular metric .

## Discussion

Using a whole-report measure of working memory confidence, we found that observers had reliable knowledge of the number of items stored on a given working memory trial. Confidence ratings, like accuracy, fluctuated from trial to trial. Overall, participants had excellent insight into the number of items stored in working memory. The number of correct items consistently correlated with the number of confident items on a trial-by-trial basis. However, resolution (correlation) and bias (over- or underconfidence) are dissociable aspects of metacognition (Koriat, 2007). While confidence and accuracy correlated, participants were particularly likely to underreport failure trials. On average, participants only correctly identified about 28% of lapse trials.

Importantly, observers' reliable metaknowledge was not an artifact of response order or temporal delay. In Experiment 3a, observers were allowed to report the items in any order they chose. Consequently, both the correct items and confident items were the first items reported in the trial. As such, observers could simply report that they were confident about the early items without having awareness of item-by-item accuracy. In Experiment 3b, we replicated this pattern for freely ordered responses, and we also added a condition where participants had to respond to the items in a randomized order. In the random order condition, response order was far less predictive of accuracy. We once again found a reliable relationship between the number of confident items and the number of correct items, although now the confident responses were distributed more equally among responses due to the random probing procedure. The random response-order condition revealed that output interference did not account for the precipitous decline in accuracy across responses in the free response-order condition. Rather, participants were aware of and chose to report their best remembered items first. When the computer forced participants to report items in a randomized fashion, the decline in performance was much less severe (7% relative to 58% from the first three to the last three responses).

Finally, we examined individual differences in the discrepancy between perceived performance (confidence) and actual performance (accuracy). Previous work has shown that low-performing individuals have particularly inflated estimates of

how their own performance compares to others' (i.e., the Dunning-Kruger effect; Kruger & Dunning, 1999), and that they also overestimate their raw performance (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008). Here, we replicate the finding that low-performing individuals overestimate their raw performance relative to high-performing individuals. There was a significant interaction between participants' quartile and misestimation of lapse rates (Experiments 3a and 3b) and mean performance (Experiment 3a only). This result was not an artifact of an extreme-groups split; underestimation of lapse rate also significantly correlated with average performance in both samples. In sum, all subjects were poor at identifying working memory failures, but those with the worst performance were doubly burdened with especially poor metacognitive awareness.

We feel it is important to point out criticisms of work related to the Dunning-Kruger effect and how those criticisms may or may not apply to our own conclusions. The main criticism of the Dunning-Kruger effect has focused on the general tendency for subjects to rate themselves as above average relative to others (Burson, Larrick, & Klayman, 2006), and how this positive bias in combination with regression to the mean could potentially explain the wider self-perception gap for low-performing individuals (Krueger & Mueller, 2002; but for counterargument, see Ehrlinger et al., 2008). Importantly, these criticisms are aimed at a particular aspect of the Dunning-Kruger model—whether metacognition truly accounts for inaccuracy of self-perception. In fact, critics of the Dunning-Kruger effect agree that there is a relationship between task-related metacognitive accuracy and task performance (Krueger & Mueller, 2002); they disagree about whether metacognitive accuracy explains the accuracy of self-perception (which we have not tested). If we were to be conservative, we should be wary that our difference score metrics might be susceptible to similar problems that have been pointed out for self-perception difference scores (namely, positive bias plus regression to the mean). Additional work is needed to assess the scope of this concern (see the Supplementary Materials for additional discussion of individual differences). Importantly, however, our trial-by-trial correlation metric is free of this criticism, as it decouples bias (intercept) from accuracy (slope); the results from our correlation metric nicely converge with our overestimation metric (perceived – actual performance), supporting our conclusion that metacognitive accuracy predicts working memory performance.

## General discussion

Across three experiments, we showed that estimates of thought content, attention state, and the number of confident representations strongly predicted working memory performance. First, we assessed the relationship between fluctuations in working memory performance and typical subjective measures of thought content. Reports of off-task thoughts (mind-wandering, task-related interference, and external distraction) all predicted a decline in working memory performance and an increased propensity for lapse trials (Experiment 1). Likewise, more continuous ratings of the degree of being "on task" covaried with fluctuations in working memory performance (Experiment 2). Second, we had participants directly report confidence for all items in all trials (Experiment 3). This whole-report confidence measure revealed a tight correspondence between the number of confident items and the number of correct items. However, this correspondence was positively biased, whereby participants were overconfident and particularly insensitive to extreme failures.

Across the board, subjective judgments were meaningfully related to performance, but participants were poor at noticing failures. In Experiments 1 and 2, participants were less likely to have a lapse in working memory performance when they reported they were focused on the task. However, a large degree of lapses persisted (5%–10%) when participants reported being "fully on task." Given baseline lapse rates of 10% to 15%, this means that the reduction in lapses for reporting "on task" was far from perfect; lapses typically went unnoticed more often than they were caught. Further, our novel measure of confidence at the item level (Experiment 3) revealed that most subjects (82%) correctly detected less than half of lapse trials. Even among the subjects who noticed more than half, sensitivity was still very poor; this "high-performing" subset of subjects still missed around 33% of lapses. While all subjects were poor at detecting lapses, some were more poor than others; subjects who performed poorly on the working memory task more greatly underestimated their failure rate.

Why might participants be unaware of working memory failures? First, when observers are in an inattentive state, they may be inattentive to both primary task demands (remembering the items) and secondary task demands (noticing which items were remembered). This possibility would be consistent with lapses where participants engage in mind-wandering and are perceptually decoupled from the task at hand (Schooler et al., 2011; Smallwood et al., 2007). Alternatively, working memory performance and metacognitive monitoring may both depend upon a common mechanism of executive control; if participants experience an executive failure, then both working memory and metacognition may suffer. A third account of overconfidence is that participants truly have some degree of information in mind (e.g., colors of squares), but they are unaware of errors in this information (e.g., binding errors). Because we asked participants to dichotomize their confidence as either "some information" about the item or "no information" about the item, some amount of the overconfidence that we observed could be attributed to trials where

participants had imprecise representations of the items that led to swap errors. Continuous-report measures of working memory (e.g., Bays & Husain, 2008; Wilken & Ma, 2004; Zhang & Luck, 2008) may be useful for measuring participants' awareness of binding errors and the degree to which feature similarity affects the rate and awareness of these errors. Of course, these accounts could all contribute to performance to varying degrees, and it will be important to disentangle the relative contribution of each.

Failures of attention and working memory are frequent, persistent, and can have devastating real-world outcomes (Reason, 1984). Here, we found that despite reliable introspection about working memory contents, observers were often insensitive to working memory failures. To close, we raise three potential avenues for future research.

### Meta-awareness of executive failures may underlie individual differences in working memory capacity

Previously, we proposed a model where fluctuations in attentional control could account for individual differences in visual working memory capacity (Adam et al., 2015). According to this model, most individuals share a common "true" visual WM capacity limit of around three simple items, and apparent individual differences in capacity are caused by how consistently individuals maximally deploy available resources. In this view, both high- and low-capacity individuals have the same potential capacity but differ dramatically in how frequently they maximize this potential. That is, effective capacity is set by the consistency of an individual's attentional control.

The results of the present experiments raise a potential alternative account of variability in working memory performance. Namely, individual differences in the consistency of metacognitive monitoring could instead explain how frequently individuals have working memory performance failures. It could be that all individuals begin to drift away from being on task at approximately the same rate but differ in how consistently they notice and correct for this drift. If this metacognitive drift correction is rapid enough, then the consequence (poor behavioral performance) will be avoided. Thus, apparent differences in behavioral outcomes could instead be explained by underlying differences in successful metacognitive monitoring.

### When is meta-awareness important for performance?

Not all studies have found a link between meta-awareness and performance. While the Dunning-Kruger effect has been shown across a wide variety of tasks, another, almost entirely separate literature, has repeatedly found no relationship between metacognitive ability and task performance (e.g., Fleming, Weil, Nagy, Dolan, & Rees, 2010; Song et al., 2011). Studies that show no relationship between metacognition and performance have yielded insights into the neural mechanisms underlying successful metacognitive monitoring (Fleming et al., 2010) and demonstrated that metacognitive monitoring can generalize across multiple tasks (Song et al., 2011) but may also have domain-specific subcomponents (Fleming, Ryu, Golfinos, & Blackmon, 2014). Additional work has shown the potential promise of using transcranial direct current stimulation (tDCS) to modulate subjective confidence (Bona & Silvanto, 2014). But it stands to reason—how useful is improving metacognitive performance if there are no behavioral consequences? While important as a causal demonstration of the role of prefrontal networks in metacognition, causal manipulations of metaknowledge would be vastly more impactful if they related to behavior.

We hypothesize that individual differences in executive control and metacognitive monitoring rely upon a common, PFC-dependent network. Therefore, the discrepancy between studies finding some versus no relationship between task performance and metacognitive monitoring may be accounted for by the degree to which the task relies upon executive control. Future experiments could take advantage of two dissociable aspects of working memory to test this hypothesis. Namely, in working memory tasks that use a continuous feature space, one can extract estimates of two components of working memory: (1) quality, or the precision with which an item is remembered, and (2) capacity, or the number of items remembered from a display. Previously, it was found that capacity predicted an important executive control ability (general fluid intelligence), but precision had no relationship with this critical ability (Fukuda, Vogel, Mayr, & Awh, 2010). As such, we predict that metacognitive ability would not predict sensory-dependent working memory precision but would, in contrast, predict executive-dependent working memory capacity. Consistent with this notion, previous studies that found no relationship between task performance and metacognitive ability typically employed low-level, sensory-dependent tasks, like perceptual monitoring (Fleming et al., 2010; Song et al., 2011) and a variant on a memory precision task (Bona & Silvanto, 2014).

### Improving meta-awareness, improving performance

Given the metacognitive blind spot toward performance failures, interventions that teach individuals to tune in to failure trials could greatly improve performance and decrease the impact of fluctuations of attention. Indeed, we recently found that feedback emphasizing failure trials was far more effective than simple feedback about performance alone (Adam & Vogel, 2016) in improving working memory performance. Future work is needed to see if such feedback benefits persist after ongoing feedback is taken away and if metacognitive sensitivity to lapses is increased during feedback. In addition

to points-based feedback, which alters subjects' intrinsic motivation (Miranda & Palmer, 2014), extrinsic motivational factors like reward may be used to improve metacognitive sensitivity to failures (Mrazek et al., 2012; Zedelius, Broadway, & Schooler, 2015). Feedback about failures could be a potentially fruitful mechanism for improving metaknowledge and overall task performance, both in the laboratory and in real-world settings. After all, eliminating failures is impossible if individuals are unaware that they have failed in the first place.

**Acknowledgements** Research was supported by NIH Grant 5R01 MH087214-08 and Office of Naval Research Grant N00014-12-1-0972. Datasets for all experiments are available online on Open Science Framework at https://osf.io/syv5w/.

**Contributions** K.A. and E.V. designed the experiments and wrote the manuscript. K.A. collected the data and performed analyses.

**Compliance with ethical standards**

**Conflicts of interest** None.

**Significance statement** Momentary failures to stay on task have consequences for ongoing task performance, from relatively minor (e.g., slow reaction time) to severe (e.g., a fatal car accident). The ability to monitor ongoing performance may be key to preventing failures. We found that subjective reports of being "on task" tracked working memory (WM) performance, but imperfectly. In particular, participants frequently reported being on task during failures. Unfortunately, on-task reports are wholly subjective, so we measured metacognitive accuracy by comparing trial-by-trial confidence to accuracy. Metacognitive accuracy predicted individual differences in WM performance, suggesting that accurate metacognitive monitoring may be key to WM success.

## References

Adam, K. C. S., Mance, I., Fukuda, K., & Vogel, E. K. (2015). The contribution of attentional lapses to individual differences in visual working memory capacity. *Journal of Cognitive Neuroscience, 27*(8), 1601–1616. doi:10.1162/jocn_a_00811

Adam, K. C. S., & Vogel, E. K. (2016). Reducing failures of working memory with performance feedback. *Psychonomic Bulletin & Review, 23*(5), 1520–1527. doi:10.3758/s13423-016-1019-4

Antrobus, J. S. (1968). Information theory and stimulus-independent thought. *British Journal of Psychology, 59*(4), 423–430. doi:10.1111/j.2044-8295.1968.tb01157.x

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321*(5890), 851–854. doi:10.1126/science.1158023

Bona, S., & Silvanto, J. (2014). Accuracy and confidence of visual short-term memory do not go hand-in-hand: Behavioral and neural dissociations. *PLoS ONE, 9*(3), e90808. doi:10.1371/journal.pone.0090808

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

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology, 90*(1), 60–77. doi:10.1037/0022-3514.90.1.60

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*(1), 98–121. doi:10.1016/j.obhdp.2007.05.002

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working memory* (pp. 102–134). Cambridge: Cambridge University Press.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain, 137*(10), 2811–2822. doi:10.1093/brain/awu221

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329*(5998), 1541–1543. doi:10.1126/science.1191883

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications, 3,* 1229. doi:10.1038/ncomms2237

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review, 17*(5), 673–679. doi:10.3758/17.5.673

Head, J., & Helton, W. S. (2016). The troubling science of neurophenomenology. *Experimental Brain Research*. doi:10.1007/s00221-016-4623-7

Huang, L. (2010). Visual working memory is better characterized as a distributed resource rather than discrete slots. *Journal of Vision, 10*(14), 8–8. doi:10.1167/10.14.8

Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin-Germeys, I., & Kwapil, T. R. (2007). For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. *Psychological Science, 18*(7), 614–621. doi:10.1111/j.1467-9280.2007.01948.x

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology, 1,* 238. doi: 10.3389/fpsyg.2010.00238

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *Cambridge handbook of consciousness* (pp. 289–325). New York: Cambridge University Press.

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality & Social Psychology, 82*(2), 180–188. doi:10.1037//0022-3514.82.2.180

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. doi:10.1037/0022-3514.77.6.1121

McKiernan, K. A., D'Angelo, B. R., Kaufman, J. N., & Binder, J. R. (2006). Interrupting the "stream of consciousness": An fMRI investigation. *NeuroImage, 29*(4), 1185–1191. doi:10.1016/j.neuroimage.2005.09.030

Miranda, A. T., & Palmer, E. M. (2014). Intrinsic motivation and attentional capture from gamelike features in a visual search task. *Behavior Research Methods, 46*(1), 159–172. doi:10.3758/s13428-013-0357-7

Mrazek, M. D., Smallwood, J., Franklin, M. S., Chin, J. M., Baird, B., & Schooler, J. W. (2012). The role of mind-wandering in measurements of general aptitude. *Journal of Experimental Psychology: General, 141*(4), 788–798. doi:10.1037/a0027968

Mutluturk, A., & Boduroglu, A. (2014). Effects of spatial configurations on the resolution of spatial working memory. *Attention, Perception,*

*& Psychophysics, 76*(8), 2276–2285. doi:10.3758/s13414-014-0713-4

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

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), 21–21. doi:10.1167/12.13.21

Reason, J. T. (1984). Lapses of attention in everyday life. In R. Paraswaman & D. R. Davies (Eds.), *Varieties of attention* (pp. 515–549). Orlando: Academic Press.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences of the United States of America, 105*(16), 5975–5979. doi:10.1073/pnas.0711295105

Schooler, J. W., Reichle, E. D., & Halpern, D. V. (2004). Zoning out while reading: Evidence for dissociations between experience and metaconsciousness. In D. T. Levin (Ed.), *Thinking and seeing: Visual metacognition in adults and children* (pp. 203–226). Cambridge: MIT Press.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences, 15*(7), 319–326. doi: 10.1016/j.tics.2011.05.006

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Dialogue: The Official Newsletter of the Society for Personality and Social Psychology, 26,* 4–7.

Smallwood, J., McSpadden, M., & Schooler, J. W. (2007). The lights are on but no one's home: Meta-awareness and the decoupling of attention when the mind wanders. *Psychonomic Bulletin & Review, 14*(3), 527–533. doi:10.3758/BF03194102

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition, 20*(4), 1787–1792. doi:10.1016/j.concog.2010.12.011

Stawarczyk, D., Majerus, S., Maj, M., Van der Linden, M., & D'Argembeau, A. (2011). Mind-wandering: Phenomenology and function as assessed with a novel experience sampling method. *Acta Psychologica, 136*(3), 370–381. doi:10.1016/j.actpsy.2011.01.002

Teasdale, J. D., Proctor, L., Lloyd, C. A., & Baddeley, A. D. (1993). Working memory and stimulus-independent thought: Effects of memory load and presentation rate. *European Journal of Cognitive Psychology, 5*(4), 417–433. doi:10.1080/09541449308520128

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology, 71,* 1–26. doi:10.1016/j.cogpsych.2014.01.003

Unsworth, N., & McMillan, B. D. (2014a). Fluctuations in pre-trial attentional state and their influence on goal neglect. *Consciousness and Cognition, 26,* 90–96. doi:10.1016/j.concog.2014.03.003

Unsworth, N., & McMillan, B. D. (2014b). Trial-to-trial fluctuations in attentional state and their relation to intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(3), 882–891. doi:10.1037/a0035544

Unsworth, N., & Robison, M. K. (2016). The influence of lapses of attention on working memory capacity. *Memory & Cognition, 44*(2), 188–196. doi:10.3758/s13421-015-0560-0

Vandenbroucke, A. R. E., Sligte, I. G., Barrett, A. B., Seth, A. K., Fahrenfort, J. J., & Lamme, V. A. F. (2014). Accurate metacognition for visual sensory memory representations. *Psychological Science, 25*(4), 861–873. doi:10.1177/0956797613516146

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12), 1120–1135. doi:10.1167/4.12.11

Zedelius, C. M., Broadway, J. M., & Schooler, J. W. (2015). Motivating meta-awareness of mind wandering: A way to catch the mind in flight? *Consciousness and Cognition, 36,* 44–53. doi:10.1016/j.concog.2015.05.016

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233–235. doi:10.1038/nature06860