Supplementary Material

For

Prospective and retrospective awareness of moment-to-moment fluctuations in visual working memory performance

Olga Kozlova[1], Kirsten C.S. Adam[1], & Keisuke Fukuda[2,3]

[1]Rice University

[2]University of Toronto Mississauga

[3]University of Toronto

Contents:

1. Simulation of prospective failure trials (Experiments 1 and 2)

Experiment 1

In our main analysis, we investigated the behaviors preceding and following the moments in which participants experienced a prospective failure (i.e., bet > number of correct reports). The observed behavioral results of Experiment 1 (see Figure 5A) suggest that participants significantly increased their bets during a prospective failure trial compared to their average bet of the three preceding trials. Moreover, participants failed to accurately adjust their total number of high confidence (HC) responses when VWM performance suffered a reduction at the prospective failure trial, though participants did indicate partial retrospective awareness. After a prospective failure trial, we found that VWM performance did not immediately recover to its pre-prospective failure baseline performance. However, participants made post-error metacognitive adjustments, such that they adjusted their following bets and HC responses to better reflect their recovering VWM performance, indicating a post-failure awareness. To investigate whether these observed results were reliable behavioral trends, we aimed to compare our observed results with a simulation of behaviors under the null hypothesis that participants had no prospective and/ or retrospective awareness of moment-to-moment VWM performance (see Figure 5B and Method).

Prospective failure trial. First, our simulation replicated the observed finding of a significant decrease in VWM performance during a prospective failure trial. Specifically, VWM performance significantly decreased during the simulated prospective failure trials (M = 1.76, SD = 0.22) compared to the simulated baseline performance on the preceding three trials (M = 2.64, SD = 0.25; $t(86)$ = 73.92, p < .001, Cohen's d = 7.93). Considering that the simulated prospective failure trials are defined as trials in which the number of correct reports happened to be lower than randomly assigned bets, this finding is expected because trials with lower number of correct reports are more likely to be classified as the simulated prospective failure trials than trials with higher number of correct reports. Interestingly however, a Kolmogorov-Smirnov test showed that the distribution of the VWM performance reduction significantly differed between the simulated results and the observed results at a prospective failure ($D(87) = 0.74$, $p < .001$). Specifically, the difference scores between the average VWM performance during the preceding baseline and the average VWM performance at a prospective failure trial were greater in the observed data ($M = 1.23$, $SD = 0.33$) compared to the simulated data ($M = 0.87$, $SD = 0.11$) Thus, in the moment of observed prospective failure trials, participants in Experiment 1 suffered a greater reduction in VWM performance than expected from the simulation.

Next, we investigated whether the observed bet increase during this prospective failure was reliably related to the drop in VWM performance. If so, our simulation should not produce a reliable bet increase on the simulated prospective failure trials. On the contrary to this prediction, our simulation produced a significant increase in participants' bets on the simulated prospective failure trial ($M = 3.16$, $SD = 0.26$) compared to their simulated pre-failure baseline bets ($M = 2.91$, $SD = 0.26$; $t(86) = 22.92$, $p < .001$, Cohen's $d = 2.46$). This suggests that the awareness of trial-by-trial performance fluctuations was not necessary to produce our observed bet increase at a prospective failure trial. In other words, placing random bets was sufficient to produce a reliable increase in bet on the simulated prospective failure trials. Interestingly, a Kolmogorov-Smirnov test showed that the distribution of the bet increase significantly differed between the simulated ($M = -0.26$, $SD = 0.11$) and the observed ($M = -0.12$, $SD = 0.15$; $D(87) = 0.51$, $p < .001$) data with a larger increase for the simulation. This result is consistent with our finding that participants had weak prospective awareness of upcoming VWM failure.

Likewise, we investigated whether the observed adjustment in HC responses in the prospective failure trials could be replicated under the null hypothesis of having no retrospective awareness. Compared to the average simulated HC responses at the pre-prospective failure baseline ($M = 2.66$, $SD = 0.53$), the HC responses at the simulated prospective failure trial were not significantly different ($M = 2.66$, $SD = 0.54$; $t(86) = 0.63$, $p = .529$, Cohen's $d = 0.07$). Thus, the simulation suggests that without any retrospective awareness, participants would not have adjusted their HC responses as observed in Figure 5A, meaning our findings of a "partial retrospective awareness" during prospective failures was a reliable trend.

Post-prospective failure trial recall performance. After the initial prospective failure, we observed a sustained reduction in VWM performance. Thus, in this simulation, we also compared the recall performance for the simulated post-failure trials to that of pre-failure trials. In contrast to our observed results, the simulated post-failure trials showed miniscule but statistically significant improvement in recall performance for three consecutive trials ($M = 2.68$, 2.68, 2.68; $SD = 0.25$, 0.25, 0.25 for +1, +2, and +3 trials, respectively) compared to the simulated pre-failure baseline performance ($M = 2.64$, $SD = 0.25$; $t(86) > 21.77$, $p < .001$, Cohen's $d > 2.33$). Thus, our simulation suggests that the sustained reduction in VWM performance following a prospective failure was a reliable observation. Importantly, however, these results do not imply whether the observed sustained reduction was a strategy or natural product of the initial drop in performance. Because we did not replicate the observed reduction in performance, we did not extend the analysis for the bets and HC responses after the prospective failure trial since such results cannot be compared to the observed trend.

Experiment 2

Prospective failure trial. For Experiment 2, we also replicated our observed finding of a significant decrease in VWM performance during a prospective failure trial. Specifically, VWM performance significantly decreased during the simulated prospective failure trials (M = 1.62, SD = 0.22) compared to the simulated baseline performance on the preceding three trials (M = 2.57, SD = 0.21; $t(84)$ = 63.01, $p < .001$, Cohen's d = 6.83). Moreover, a Kolmogorov-Smirnov test showed that the distribution of the VWM performance reduction significantly differed between the simulated results and the observed results at a prospective failure ($D(85)$ = 0.62, $p < .001$). Specifically, the difference scores between the average VWM performance during the preceding baseline and the average VWM performance at a prospective failure trial were greater in the observed data (M = 1.17, SD = 0.74) compared to the simulated data (M = 0.95, SD = 0.14) Thus, in the moment of observed prospective failure trials, participants in Experiment 2 suffered a greater reduction in VWM performance than expected from the simulation.

Next, our simulation replicated the same slight, but reliable, increase in participants' bets during the prospective failure trial (M = 2.98, SD = 0.28) compared to their simulated pre-failure baseline bets (M = 2.67, SD = 0.30; $t(84)$ = 23.59, $p < .001$, Cohen's d = 2.56). This suggests that no additional mechanism– other than having no prospective awareness– was necessary to produce our observed bet increase at a prospective failure trial in Experiment 2 as well. Interestingly, a Kolmogorov-Smirnov test showed that the distribution of the bet increase significantly differed between the simulated (M = -0.31, SD = 0.12) and the observed (M = -0.19, SD = 0.19; $D(85)$ = 0.46, $p < .001$) data with a larger increase for the simulation. This result is consistent with our finding that participants had weak prospective awareness of upcoming VWM failure. Lastly, we found that compared to the average simulated HC responses at the pre-prospective failure baseline (M = 2.61, SD = 0.38), the HC responses at the simulated prospective failure trial were not significantly different ($M = 2.61$, $SD = 0.38$; $t(84) = 0.41$, $p = .686$, Cohen's $d = 0.04$). Thus, the simulation suggests that without any retrospective awareness, participants would not have adjusted their HC responses as observed in Figure 10A, meaning our findings of a "partial retrospective awareness" during prospective failures was a reliable trend. Altogether, our results suggest that participants were again randomly placing bets during the prospective failure trial to give rise to the observed bet increase, though their observed HC responses during the prospective failure trials were reflective of reliable retrospective metacognition.

Post-prospective failure trial recall performance. In contrast to our observed results, the simulated post-failure trials in Experiment 2 also showed miniscule but statistically significant improvement in recall performance for three consecutive trials ($M = 2.62$, 2.62, 2.62; $SD = 0.22$, 0.22, 0.22 for +1, +2, and +3 trials, respectively) compared to the simulated pre-failure baseline performance ($M = 2.57$, $SD = 0.21$; $t(84) > 20.52$, $p < .001$, Cohen's $d > 2.23$). Thus, our simulation suggests that the sustained reduction in VWM performance following a prospective failure was a reliable observation. As in Experiment 1, since we did not replicate the observed reduction in performance, we did not extend the analysis for the bets and HC responses after the prospective failure trial since such results cannot be compared to the observed trend.

2. Individual differences during and surrounding prospective failures (Experiments 1 and 2)

Experiment 1

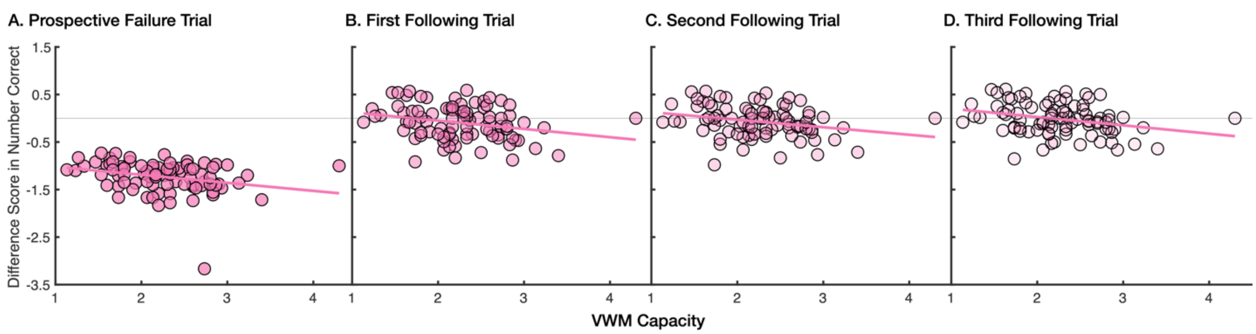
High-capacity individuals recover more slowly from reductions in VWM performance. In our main analysis, we investigated how VWM performance was affected during a prospective failure. There, we found that individuals suffered a reduction in VWM performance compared to their pre-prospective failure baseline trials. Did high- and low-capacity individuals suffer an equal reduction in performance? To investigate the individual differences in performance reductions, we calculated the difference score between each participant's average VWM performance during a prospective failure and their respective average pre-failure baseline performance. Figure S1A suggests that all individuals suffered a reduction in performance. However, we found that high-capacity individuals experienced a greater reduction in VWM performance during a VWM failure compared to low-capacity individuals ($r(85) = -.27$, $p < .05$). Thus, high-capacity individuals appear to be affected by the prospective failure at a greater magnitude, likely due to them correctly reporting more items than low-capacity individuals before the prospective failure occurred.

In addition to the initial reduction of VWM performance during a prospective failure trial, we found that VWM performance did not recover to pre-failure baseline performance until the third following trial (see Figure 5). However, it is possible that individuals differed in their performance recovery, along with the post-failure awareness. To investigate individual differences in the effects of prospective failures on following performance, we similarly calculated the difference score between each individual's first, second, and third, following trial and their pre-failure baseline. Right after the prospective failure, low-capacity individuals appeared to recover at a greater rate than high-capacity individuals ($r(85) = -.26$, $p < .05$; see Figure S1B). Moreover, this rate of performance recovery for low-capacity individuals extended to the second ($r(85) =$

-.26, $p < .05$; see Figure S1C), and third following trial ($r(85) = -.29$, $p < .01$; see Figure S1D). Altogether, individuals differed in the impact of a prospective failure when it occurred, but high-capacity individuals were slower at recovering to their pre-failure baseline performance compared to low-capacity individuals in the following trials.

Figure S1

Experiment 1: Differences in Performance at Prospective Failure and Post-Failure

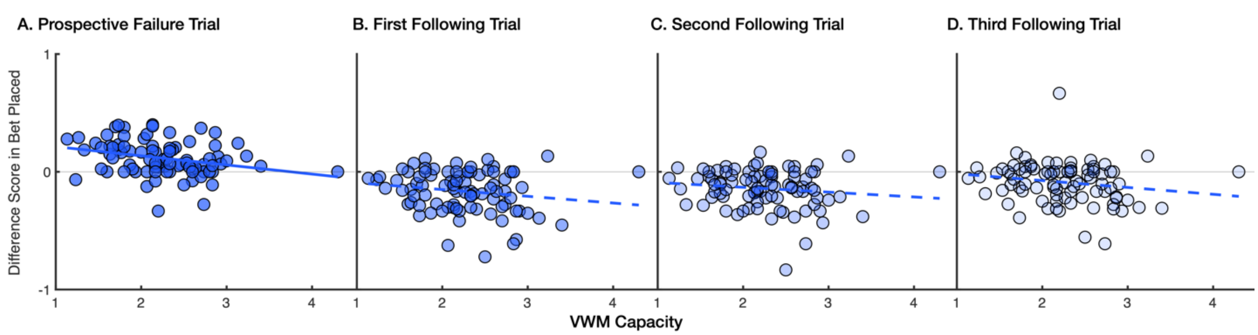


Note. $N = 87$. To denote when participants did not differ between pre-prospective failure performance and the trial of interest's performance, the figures display a solid line along the graph's origin. Solid trend lines indicate a significant correlation. The change in the number of correctly reported colors was determined by subtracting each individual's pre-prospective failure baseline recall performance from the trial of interest's performance (e.g., prospective failure performance – pre-failure baseline performance). VWM capacity scores were estimated from each individual's performance in the whole-report baseline task. (A) Difference score between the prospective failure trial and pre-failure baseline performance. (B) Difference score between the first following trial of a prospective failure and pre-failure baseline performance. (C) Difference score between the second following trial of a prospective failure and pre-failure baseline performance. (D) Difference score between the third following trial of a prospective failure and pre-failure baseline performance.

Individual differences in bets during a prospective failure, but similar post-failure bet adjustments. Could high-capacity individuals more accurately predict the upcoming reduction in performance than low-capacity individuals? To investigate individual differences in predictions at prospective failures trials, we similarly calculated the difference between each individual's bet during a prospective failure and their pre-prospective failure baseline bet value. Here, we found that low-capacity individuals increased their bets compared to baseline bets more than high-capacity individuals ($r(85) = -.29$, $p < .01$). Thus, low-capacity individuals appear to drive the overall trend in increasing bets at the failure trial (see Figure S2A). Were individuals equally aware of their reduced VWM performance in their post-failure trials? Figures S2B, S2C, and S2D suggest that a majority of individuals lowered their bets in the following first, second, and third trials, respectively. However, we found no significant correlations between VWM capacity and the first ($r(85) = -.17$, $p = .110$), second ($r(85) = -.13$, $p = .214$), and third ($r(85) = -.18$, $p = .092$), following trial bet difference scores. Thus, following prospective failures, high- and low-capacity individuals were equally likely to lower their bets to accommodate the general trend in suffering a reduction of performance. Interestingly, however, when these results are paired with the trends in performance recovery (see Figure S1), they may suggest that high-capacity individuals were employing a strategy of reduced bets and recall performance to avoid any further losses of points. Specifically, as illustrated in Figure S2B and beyond, low-capacity individuals trend towards adopting their pre-failure baseline bets at a quicker rate than high-capacity individuals.

Figure S2

Experiment 1: Differences in Prospective Awareness at Prospective Failure and Post-Failure



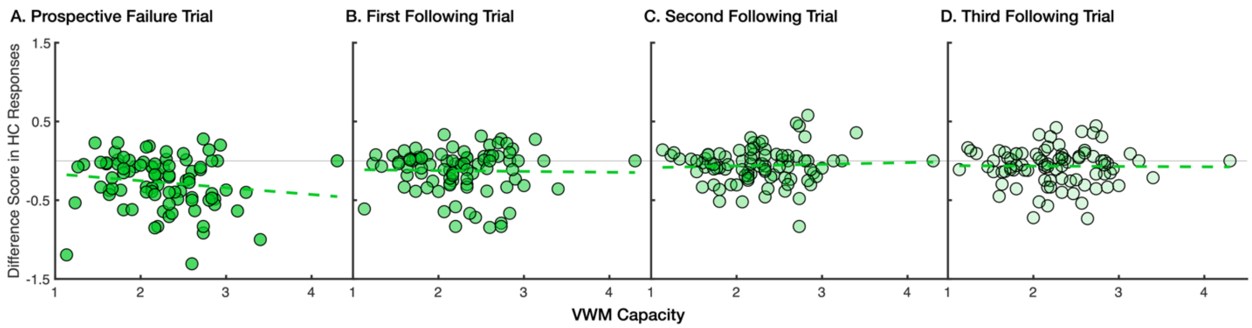
Note. $N = 87$. Solid lines indicate a significant correlation. The difference score in bets was determined by subtracting each individual's pre-prospective failure baseline bet placed from the trial of interest's bet (e.g., prospective failure bet – pre-failure baseline bet). (A) Difference score between the prospective failure trial and pre-failure baseline bets. (B) Difference score between the first following trial of a prospective failure and pre-failure baseline bets. (C) Difference score between the second following trial of a prospective failure and pre-failure baseline bets. (D)

Difference score between the third following trial of a prospective failure and pre-failure baseline bets.

No capacity-related differences in retrospective awareness during and after a prospective failure. If low-capacity individuals had poor prospective awareness of the upcoming reduction in performance, did they also fail to decrease their high confidence (HC) at a prospective failure trial? Just as before, we calculated the difference between the total number of HC responses during a prospective failure compared to each individual's pre-prospective failure baseline HC responses. We found no significant correlation between the magnitude of the changed HC responses and VWM capacity during a prospective failure trial ($r(85) = -.14$, $p = .186$). Additionally, there was no correlation between VWM capacity and the total number of HC responses during a prospective failure ($r(85) = .14$, $p = .195$). However, Figure S3 suggests that individuals tended to decrease their HC responses, signified by majority of the difference scores being below zero. Thus, high- and low-capacity individuals reflected on the VWM failure to similar extents. Similarly, since we observed a reduction in the HC responses in the following three trials of a prospective failure, we investigated whether VWM capacity scores predicted this reduction. If low-capacity individuals are more likely to be overall overconfident in their performance, then perhaps they do not reduce their confidence responses as much as high-capacity individuals. However, we found no significant correlation between the difference scores in HC responses and VWM capacity in the first ($r(85) = -.02$, $p = .844$), second ($r(85) = .05$, $p = .649$), and third ($r(85) = -.01$, $p = .922$) following trials of a prospective failure (see Figures S3B, S3C, and S3D, respectively). Thus, high- and low-capacity individuals reduced their confidence in similar fashions during the post-failure trials

Figure S3

Experiment 1: Differences in Retrospective Awareness During and After Prospective Failures



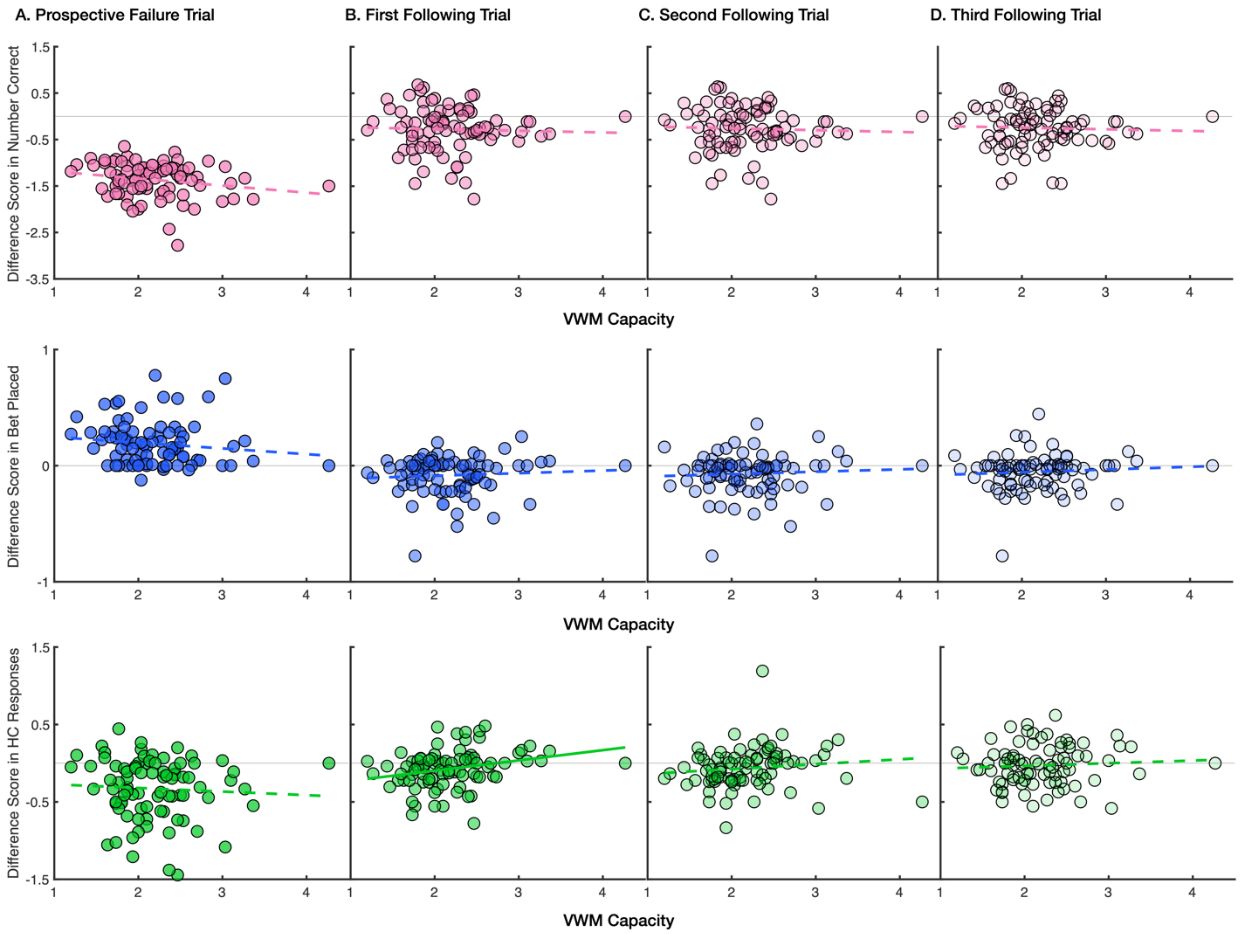
Note. N = 87. Solid lines indicate a significant correlation. The change in the total number of high confidence responses was determined by subtracting each individual's pre-failure baseline HC responses from the trial of interest's HC responses (e.g., total number of high confidence responses at a prospective failure– pre-failure baseline total number of HC responses). (A) Difference score between the prospective failure trial and pre-failure baseline HC responses. (B) Difference score between the first following trial of a prospective failure and pre-failure baseline HC responses. (C) Difference score between the second following trial of a prospective failure and pre-failure baseline HC responses. (D) Difference score between the third following trial of a prospective failure and pre-failure baseline HC responses.

Experiment 2

In Experiment 2, participants did not receive trial-by-trial feedback of their VWM performance. There, our main analysis replicated our findings from Experiment 1, such that VWM performance suffered a reduction during a prospective failure trial, and participants failed to adjust their bets and HC responses to reflect the full severity of their reduced performance. However, perhaps the absence of feedback modulated the capacity-related individual differences that were observed in Experiment 1. More specifically, we predicted that the previously observed capacity-related individual differences may be amplified, such that low-capacity individuals would experience more severe metacognitive blindness than high-capacity individuals during a prospective failure trial along with the post-failure trials. Thus, we similarly calculated the difference score for each participant between the pre-prospective failure behaviors and the trial of interest's behaviors (see Figure S4).

Figure S4

Experiment 2: Differences in Performance and Awareness at Prospective Failure and Post-Failure



Note. $N = 85$. Solid lines indicate a significant correlation. The difference scores were determined by subtracting each individual's pre-failure baseline response from the trial of interest's high response (e.g., VWM performance at a prospective failure trial– pre-failure baseline VWM performance). (A) Difference scores between the prospective failure trial and pre-failure baseline responses. (B) Difference score between the first following trial of a prospective failure and pre-failure baseline. (C) Difference score between the second following trial of a prospective failure and pre-failure baseline. (D) Difference score between the third following trial of a prospective failure and pre-failure baseline.

No capacity-related differences in performance during or after prospective failures. At the prospective failure trial, we found a similar trend as Experiment 1, such that high-capacity individuals suffered a greater reduction in performance than low-capacity individuals (see Figure S4A). However, this correlation between VWM capacity and the magnitude of reduced performance fell short of significance ($r(83) = -.20$, $p = .062$). After the prospective failure in Experiment 2, we found that VWM performance was still reduced in the first following trial (see Figure 10). However, unlike Experiment 1, VWM capacity did not predict this performance recovery in the first following trial ($r(83) = -.04$, $p = .734$; see Figure S4B), second following trial ($r(83) = -.04$, $p = .710$; see Figure S4C), nor third trials after a prospective failure ($r(83) = -.04$, $p = .714$; see Figure S4D). Thus, the absence of feedback may have changed participants' performance recovery likely due to the adoption of a conservative strategy.

Similar bet adjustments during and after prospective failures. Were low-capacity individuals still more likely to increase their bets during a prospective failure trial, even in the absence of feedback? Figure S4 suggests that majority of individuals increased their bet at a prospective failure. However, we found no correlation between VWM capacity and the bet difference score at a prospective failure ($r(83) = -0.13$, $p = .237$; see Figure S4A), meaning high- and low-capacity individuals were equally likely to increase their bet. Did individuals equally adjust their following bets, even in the absence of feedback? We found no correlations between VWM capacity and the following first trial bet difference scores, ($r(83) = .07$, $p = .531$; see Figure S4B). Although performance had recovered by the second following trial, individuals appeared to have overshot their metacognitive adjustment to the performance recovery through their bets. Specifically, participants continued to decrease their bets in the following second and third trials of a prospective failure (see Figures S4C and S4D). During these trials, we found no significant correlations between VWM capacity and the adjusted bets in the second following trial ($r(83) = .06$, $p = .582$), nor third following trial of a prospective failure ($r(83) = .08$, $p = .485$). Thus, both high- and low-capacity individuals continued to decrease their bets after a prospective failure.

Faster retrospective awareness recovery for high-capacity individuals. The absence of feedback may have exacerbated overconfidence for low-capacity individuals in Experiment 2. However, we found no correlation between the difference scores in HC responses and VWM capacity during a prospective failure trial ($r(83) = -.06$, $p = .588$). Additionally, there was no correlation between VWM capacity and the total number of HC responses during a prospective failure ($r(83) = .14$, $p = .205$). Thus, high- and low-capacity individuals were equally sensitive in reflecting on the prospective failure. Interestingly, however, we found a correlation between VWM capacity and the following first trial difference scores in HC responses ($r(83) = .27$, $p = .013$). Specifically, Figure S4B suggests that low-capacity individuals decreased their HC responses, while high-capacity individuals approached, or went above their pre-prospective failure HC responses. High-capacity individuals may have believed VWM performance had recovered, while low-capacity individuals continued to reflect with a "reduced-performance mindset". In the following second trial, individuals did not differ in their adjusted HC responses ($r(83) = .12$, $p = .278$), though they had yet to recover to their pre-failure HC responses overall (see Figure S4C). We also found no correlation between VWM capacity and the adjusted HC responses in the third following trial ($r(83) = .07$ $p = .549$), though HC responses had recovered to their pre-prospective failure level (see Figure S4D). Thus, in the absence of feedback, high-capacity individuals may have approached their baseline HC responses at a faster rate than low-capacity individuals.

3. Remote versus in-person comparison of metacognitive and recall performance (Experiment 1)

In our main analysis, we found that participants in Experiment 2 committed fewer prospective failures than participants in Experiment 1. Moreover, our results suggest that the absence of feedback may have contributed to more conservative metacognitive and recall performance in Experiment 2 than in Experiment 1. However, it is possible that this difference may have been influenced by the in-person monitoring of Experiment 2, compared to majority of Experiment 1 conducted remotely. To address this concern, we outline a series of independent *t*-tests between the bets, total number of HC responses, and number of correct reports between the remote (n = 57) and in-person participants (n = 30) from Experiment 1. In Experiment 1, the bets placed by remote (M = 3.06, SD = 0.56) and in-person (M = 2.91, SD = 0.53) participants did not significantly differ ($t(85)$ = 1.20, p = .234, Cohen's d = 0.27). Likewise, the total number of HC responses for remote (M = 2.86, SD = 0.82) and in-person participants (M = 2.82, SD = 0.57) did not significantly differ ($t(85)$ = 0.22, p = .824, Cohen's d = 0.05). Lastly, there was no significant different in the number of correct reports between remote (M = 2.97, SD = 0.51) and in-person participants (M = 2.86, SD = 0.41; $t(85)$ = 1.00, p =.321, Cohen's d = 0.23). Welch's *t*-tests did not show any significant differences between remote and in-person participants for all three response types as well. Overall, these results suggest that there were no differences in metacognitive and/ or recall performance between remote or in-person participants.

4. Results of set-size six whole-report bet task (Experiment S1)

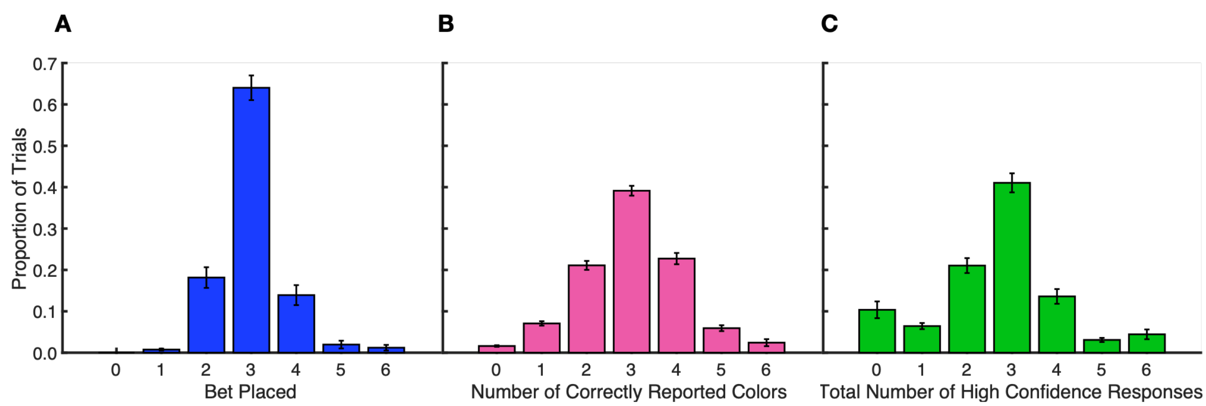
Experiment S1

In Experiments 1 and 2, we found that individuals overestimated their upcoming VWM performance for almost a quarter of the experiment. We termed these overestimation trials as "prospective failures" and investigated the severity in the misalignment between participants' bets and VWM performance in those moments. Then, we found that VWM performance suffered a reduction during these prospective failure trials, and that participants were not simply increasing their bets during stable VWM performance. Lastly, during this prospective failure and reduction in VWM performance, we found that individuals did not adjust their retrospective confidence ratings to reflect the severity of the reduced performance. However, these experiments contained memory arrays of four items, whereas individuals are exposed to a much greater amount of information in the real world. Moreover, it is possible that our set-size of four items limited the possibility of high-capacity individuals overestimating their upcoming performance or reporting overconfidence. Thus, we ran an additional experiment to investigate whether our results are replicable with a larger memory array. In Experiment S1, a unique set of participants (N = 89, mean age = 19.55, 63 women, 25 men, one chose not to report gender) participated in Experiment S1. The methods and procedures were identical to Experiment 1, except that participants saw six items in the memory array of the whole-report bet task, rather than four.

Retrospective awareness more sensitive to performance. As in Experiments 1 and 2, we also calculated the proportion of bets placed, the proportion of the total number of high confidence (HC) responses submitted, and the proportion of correctly recalled items throughout the whole-report bet task (see Figure S5). In Experiment S1, participants correctly reported 3.02 items (SD = 0.50) in the whole-report bet task. However, the average bet placed was 3.02 items (SD = 0.51), which was not statistically different than the average number of correctly reported items (t(88) = 0.01, p = .994, Cohen's d < 0.01). On the other hand, the average number of HC responses was 2.68 HC responses (SD = 0.93), and it was statistically lower than the number of correctly reported items (t(88) = 3.24, p < .01, Cohen's d = 0.34). Overall, these results suggest that participants were more likely to overestimate their VWM performance prospectively, but not retrospectively.

Figure S5

Response Proportions for Whole-Report Bet Task in Experiment S1

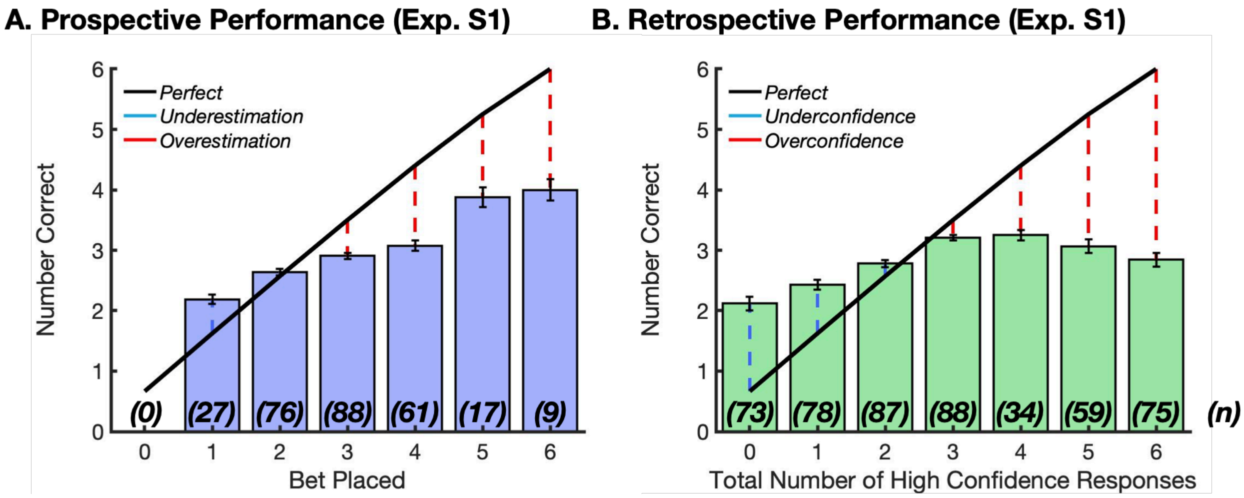


Note. Error bars represent the standard errors of the mean. (A) The proportion of trials from the whole-report bet task where participants placed a particular bet. (B) The proportion of correctly reported items from the whole-report bet task. (C) The proportion of total high confidence responses within a trial.

Simulation of accurate prospective and retrospective awareness. For Experiment S1, we also employed a Monte Carlo simulation to estimate the number of correct reports when participants' metacognitive ratings were perfect (see Method for details). The results of the metacognitive responses Monte Carlo simulation for Experiment S1 are illustrated in Figure S6.

Figure S6

Monte Carlo Simulation for Predictions and Reflections in Experiment 1S

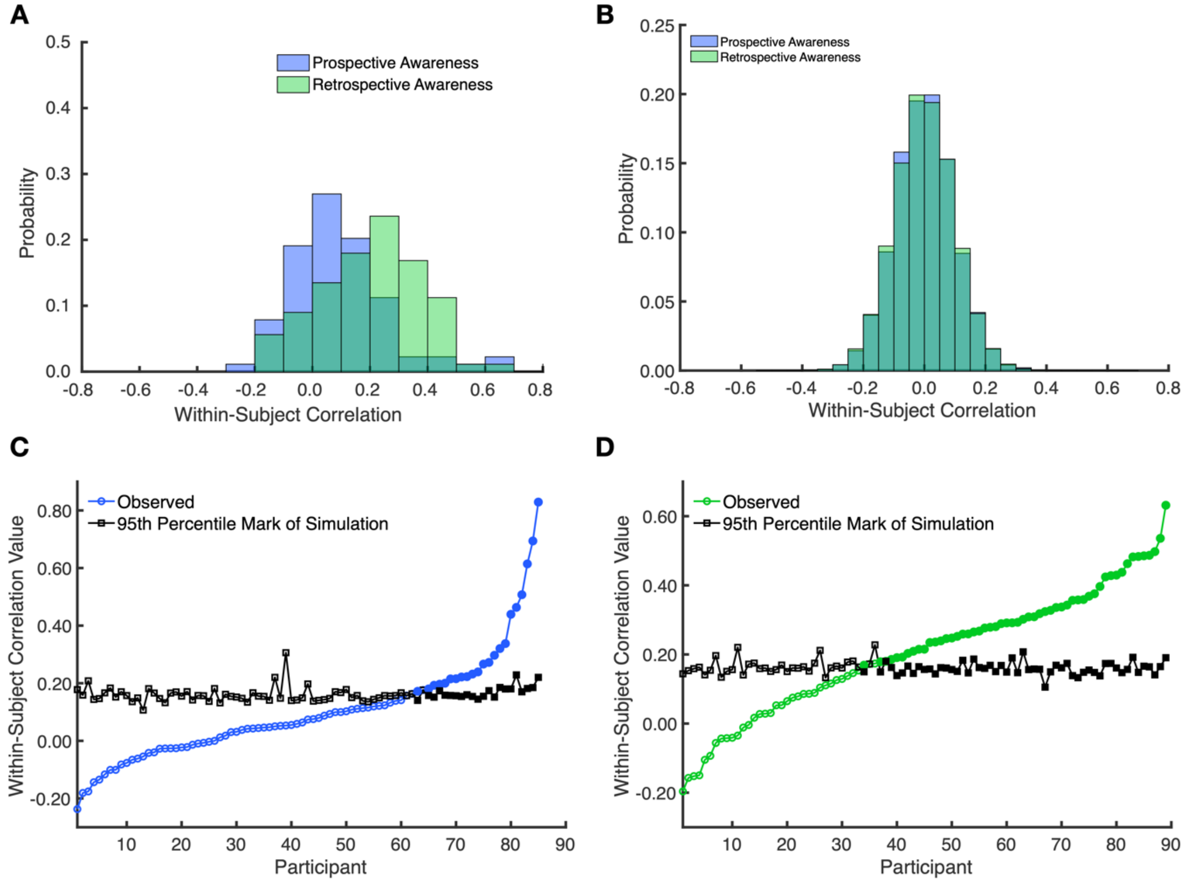


Note. N = 89. The qualitative relationship between the simulated and experimental recall performance based on each prospective and retrospective metacognitive rating. The number of participants that contributed towards a particular bet or total number of high confidence responses within a trial is denoted by n. Error bars represent the standard errors of the mean. (A) Prospective performance in Experiment S1. For each possible bet (0– 6), our Monte Carlo simulation predicted a recall performance (solid trend line). If participants had accurate prospective awareness of their upcoming performance, the bar graphs would align to the solid black line. However, participants underestimated (i.e., correct reports > bet placed) or overestimated (i.e., correct reports < bet placed) their upcoming performance as denoted by the dotted lines. (B) Retrospective performance in Experiment S1. For each possible total number of high confidence responses (0– 6), a predicted recall performance was simulated via the same Monte Carlo simulation. If participants had accurate retrospective awareness of their performance, the bar graphs would align to the solid black line. However, participants were underconfident (i.e., correct recall > total number of high confidence responses) or overconfident (i.e., correct recall < total number of HC responses) in their performance.

Prospective awareness worse than retrospective awareness. To investigate the sensitivity of prospective and retrospective judgments to VWM performance, we computed the within-subject correlations in Experiment S1 as well (see Figure S7A). The average correlation value for prospective awareness (bets) was reliably above zero (mean $r = .10$, $SD = 0.18$; $t(84) = 5.13$, $p < .001$, Cohen's $d = 0.56$). Likewise, the average correlation value for retrospective awareness (HC responses) was also reliably above zero (mean $r = .21$, $SD = 0.18$; $t(88) = 10.94$, $p < .001$, Cohen's $d = 1.16$), which was statistically higher than the prospective awareness correlation ($t(84) = 3.96$, $p < .001$, Cohen's $d = 0.44$). Taken together, we replicated our previous results, such that both prospective and retrospective awareness of VWM fluctuations were not random, and that retrospective awareness was more accurately tuned to the fluctuations of VWM performance. However, we similarly found overall low correlations, suggesting a limited extent of participants' metacognitive awareness in their moment-to-moment VWM performance.

Figure S7.

Observed and Simulated Within-Subject Correlations in Experiment 1in Experiment S1



Note. (A) The distribution of observed within-subject correlations (WSC) for prospective and retrospective awareness. To determine the WSCs for prospective and retrospective awareness, we computed the correlation between each participant's bet and total number of high confidence (HC) responses with their respective VWM performance on a trial-by-trial basis. Then, we calculated the probability distribution of WSCs. (B) The distribution of simulated WSCs. To simulate the WSCs, we conducted 10,000 permutation tests within each subject, where we shuffled each participant's number of correct reports, bets, and total number of HC responses throughout the experiment. For each of the 10,000 permutations, we calculated a WSC between the shuffled number of correct reports with the bets placed (prospective awareness) and the total number of HC responses (retrospective awareness). (C) Observed prospective awareness (WSCs) in Experiment S1 and the values representing the 95th percentile of simulated WSCs for prospective awareness. Participants with an observed WSC greater than the 95th percentile mark (colored-in data points) indicate reliable moment-to-moment prospective awareness ($p < .05$). (D) Observed retrospective awareness (WSCs) in Experiment S1 and the values representing the 95th percentile of simulated WSCs for retrospective awareness. Participants with an observed WSC greater than the 95th percentile mark (colored-in data points) indicate reliable moment-to-moment retrospective awareness ($p < .05$).

Observed group-level awareness of VWM performance is a reliable trend. To validate the weak prospective and retrospective awareness in Experiment S1, we compared our observed within-subject correlations to a simulated null hypothesis that participants in this experiment had no prospective and/ or retrospective awareness (see Method). The results of the simulated within-subject correlations (WSCs) are illustrated in Figure S7B. Here, we replicated our findings from Experiments 1 and 2, such that a null hypothesis in having no prospective and/ or retrospective awareness resulted in a distribution of WSCs that clustered around zero. We found that the average simulated WSC for prospective awareness was approximately zero and that it was not statistically different from zero (mean $r < .01$, $SD < 0.01$; $t(84) = 0.20$, $p = .845$, Cohen's $d = 0.02$). Similarly, the average simulated WSC for retrospective awareness was approximately zero and was not statistically different from zero (mean $r < .01$, $SD < 0.01$; $t(88) = 0.50$, $p = .621$, Cohen's $d = 0.05$). These results suggest that participants with no prospective and/ or retrospective awareness would have an observed WSC around zero, meaning their metacognitive ratings would not have predicted their number of correct reports. Additionally, a Kolmogorov-Smirnov test found that our observed distribution of WSCs for prospective awareness was significantly different from the simulated distribution of WSCs ($D(89) = 0.30$, $p < .001$). Thus, at the group level, participants in Experiment S1 appear to have had reliable prospective awareness. Similarly, the distribution of observed WSCs for retrospective awareness was significantly different from the simulated distribution of retrospective WSCs, meaning there was reliable moment-to-moment retrospective awareness at a group level ($D(89) = 0.59$, $p < .001$). Next, we compared the simulated WSCs to the observed WSCs to determine which individual participants had reliable moment-to-moment metacognitive awareness. Figures S7C and S7D allow for a direct comparison between the observed WSC and the simulated WSC representing whether the participant truly had moment-to-moment metacognitive awareness. We found that 22 out of 89 participants had an observed prospective WSC greater than their 95th percentile value of simulated WSCs ($p < .05$). This means that majority of participants could not reliably predict their upcoming VWM performance on a moment-to-moment basis in this experiment. On the other hand, 54 out of 89 participants had an observed retrospective WSC greater than their 95th percentile value of simulated WSCs, indicating that a majority of participants could reliably reflect on the number of items they correctly reported from moment to moment. In conclusion, majority of participants did not have reliable moment to moment prospective awareness in Experiment S1, though retrospective awareness was more calibrated from moment to moment.

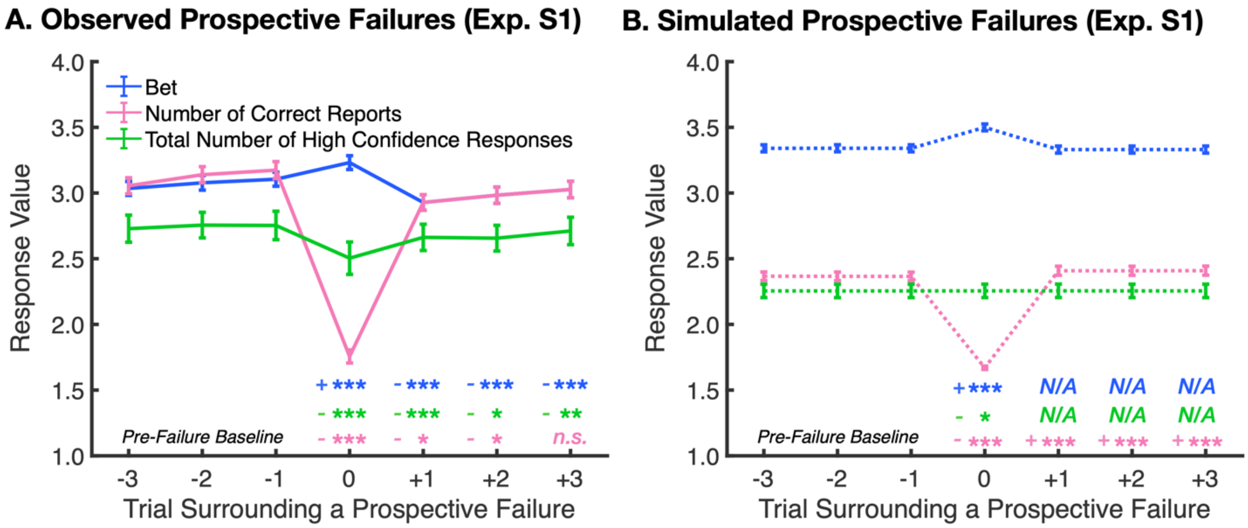
Prospective failures correspond to reductions in VWM performance. In our analysis, we defined a prospective failure as when participants placed a bet that was greater than the number of items correctly reported (i.e., they overestimated upcoming performance and earned zero points). To investigate performance surrounding a prospective failure, we similarly restricted our analyses to prospective failure trials that were preceded and followed by three trials within the same experimental block. This resulted in 24% (*SD* = 10%) of the prospective failure trials being excluded in Experiment S1– a mean of six trials across participants. Similarly, we also tracked the occurrence of complete lapses in VWM performance (0-1 correct reports) and restricted the analysis to lapse trials that were preceded and followed by three trials. This resulted in 25% (*SD* = 21%) of lapse trials being excluded in Experiment S1– a mean of two trials across participants. However, three participants from Experiment S1 were excluded from the lapse analysis due to not experiencing any lapses in performance.

About 25% of participants' total trials consisted of prospective failures (*SD* = 11%). Next, we set to identify the severity of these prospective failures– how often did individuals fail to predict a complete lapse in performance versus underperforming their bets? Within these prospective failure trials, 35% (*SD* = 18%) were associated with complete lapses in VWM performance. Thus, most prospective failure trials (*M* = 65%, *SD* = 18%) were moments in which participants correctly reported fewer colors than their bets but were not complete lapses in VWM performance. However, we found that almost all of individuals' complete lapse trials in Experiment S1 were successfully captured by our definition of a prospective failure (*M* = 99%, *SD* = 7%). Thus, individuals often failed to predict an upcoming, complete lapse in VWM performance. Overall, these lapse trials accounted for 9% (*SD* = 6%) of the experiment trials.

During a prospective failure, we replicated our previous findings, such that VWM performance significantly decreased (*M* = 1.76, *SD* = 0.47) compared to pre-prospective failure baseline performance (*M* = 3.12, *SD* = 0.55; $t(88) = 29.57$, $p < .001$, Cohen's $d = 3.13$; see Figure S8). Moreover, we replicated the same slight, but reliable, increase in participants' bets (*M* = 3.23, *SD* = 0.50) at the prospective failure trials compared to the pre-failure baseline bets (*M* = 3.07, *SD* = 0.50; $t(88) = 8.44$, $p < .001$, Cohen's $d = 0.89$). Thus, individuals were unable to predict the upcoming reduction in VWM performance.

Figure S8

Observed and Simulated Performance Around Prospective Failures in Experiment S1



Note. N = 89. The x-axis '0' represents a prospective failure trial, with its three preceding and following trials from the left to right side of the x-axis, respectively. Asterisks indicate a significant change in behaviors compared to the averaged pre-failure baseline responses (from top to bottom: bets, total number of high confidence responses, number of correct reports). The direction of the significant change is indicated by a + (current trial response > pre-failure baseline response) or – (current trial response < pre-failure baseline response). Error bars represent standard errors of the mean. The response value corresponds to the mean bet placed, mean number of correctly reported items, and mean total number of high confidence responses placed during a particular trial. (A) The observed performance surrounding a prospective failure trial in Experiment S1. (B) The simulated performance surrounding the simulated prospective failure trials for Experiment S1. Because of the slight, but significant boost in VWM performance in the following trials of a prospective failure in the simulation, the first, second and third following trials' simulated responses were uninterpretable (N/A).

*** $p < .001$
** $p < .01$
* $p < .05$

Partial recognition of a prospective failure. Could individuals reflect on their reduced performance at a prospective failure trial? In Experiment S1, participants also partially reduced their HC responses during the failure trials (M = 2.50, SD = 1.16) compared to their baseline HC responses (M = 2.75, SD = 0.95; $t(88) = 5.69$, $p < .001$, Cohen's d = 0.60). However, while participants reduced their HC in their performance, this reduction in HC responses did not capture the full extent of their reduced performance. Given the replication of partial awareness of a prospective failure, we ran another post-hoc analysis to identify how often participants were aware or unaware of their reduced VWM performance. A large majority (87%) of participants in Experiment S1 experienced both types of failure trials, except for six individuals that were always unaware, and six individuals that were consistently aware of the failures. When examining all individuals, participants were more likely to be unaware of their failure (M = 59%, SD = 30%) than aware (M = 41%, SD = 30%; $t(88) = 2.71$, $p < .01$, Cohen's d = 0.23). Even when participants experienced lapses in VWM performance, participants were more likely to be unaware (M = 66%, SD = 37%) than aware (M = 34%, SD = 37%; $t(85) = 3.96$, $p < .001$, Cohen's d = 0.43). Thus, we replicated our previous findings such that prospective failures– no matter their severity, often go undetected when they occur.

Awareness of VWM performance increases post-failure. Right after the prospective failure, recall performance was still reduced (M = 2.93, SD = 0.56) compared to pre-prospective failure baseline performance (M = 3.12, SD = 0.55; $t(88) = 3.76$, $p < .001$, Cohen's d = 0.40). Moreover, VWM performance did not recover to baseline performance in the following second (M = 2.98, SD = 0.59; $t(88) = 2.75$, $p < .01$, Cohen's d = 0.29) trial. However, VWM performance recovered by the third following trials (M = 3.03, SD = 0.60; $t(88) = 1.88$, $p = .064$, Cohen's d = 0.20). Thus, VWM performance suffered a sustained reduction for two following trials. However, participants adjusted their following bets to accommodate this reduced performance by reducing their following bet (M = 2.93, SD = 0.56) compared to baseline bets (M = 3.07, SD = 0.50; $t(88)$ = 7.24, p < .001, Cohen's d = 0.77). This trend of decreased bets continued for the second following trials (M = 2.98, SD = 0.59; $t(88)$ = 4.71, p < .001, Cohen's d = 0.50) and third following trials (M = 3.03, SD = 0.60; $t(88)$ = 2.07, p = .041, Cohen's d = 0.22). Thus, participants reduced their bets for an additional trial after VWM performance recovered, then adopted their pre-failure baseline betting behaviors. Mirroring the trajectory of the following bets, participants also reduced their HC responses in the first (M = 2.66, SD = 0.94; $t(88)$ = 3.28, p < .01, Cohen's d = 0.35), and second (M = 2.66, SD = 0.92; $t(88)$ = 3.42, p < .001, Cohen's d = 0.36), but not third (M = 2.721 SD = 0.98; $t(88)$ = 1.05, p = .295, Cohen's d = 0.11) following trials of a prospective failure compared to their pre-failure baseline HC responses (M = 2.75, SD = 0.95).

Simulation of prospective failure trials. To simulation replicated the observed finding of a significant decrease in VWM performance during a prospective failure trial. Specifically, VWM performance significantly decreased during the simulated prospective failure trials (M = 1.67, SD = 0.12) compared to the simulated baseline performance on the preceding three trials (M = 2.37, SD = 0.31; $t(88)$ = 27.40, p < .001, Cohen's d = 2.90). Considering that the simulated prospective failure trials are defined as trials in which the number of correct reports happened to be lower than randomly assigned bets, this finding is expected because trials with lower number of correct reports are more likely to be classified as the simulated prospective failure trials than trials with higher number of correct reports. Interestingly however, a Kolmogorov-Smirnov test showed that the distribution of the VWM performance reduction significantly differed between the simulated results and the observed results at a prospective failure ($D(89)$ = 0.80, p < .001). Specifically, the difference scores between the average VWM performance during the preceding baseline and the average VWM performance at a prospective failure trial were greater in the observed data (*M* = 1.37, *SD* = 0.44) compared to the simulated data (*M* = 0.70, *SD* = 0.24) Thus, in the moment of observed prospective failure trials, participants in Experiment S1 suffered a greater reduction in VWM performance than expected from the simulation.

Next, we investigated whether the observed bet increase during this prospective failure was reliably related to the drop in VWM performance. Our simulation produced a significant increase in participants' bets on the simulated prospective failure trial (*M* = 3.50, *SD* = 0.24) compared to their simulated pre-failure baseline bets (*M* = 3.34, *SD* = 0.25; $t(88) = 24.09$, $p < .001$, Cohen's *d* = 2.55). This suggests that the awareness of trial-by-trial performance fluctuations was not necessary to produce our observed bet increase at a prospective failure trial. In other words, placing random bets was sufficient to produce a reliable increase in bet on the simulated prospective failure trials. Interestingly, a Kolmogorov-Smirnov test showed that the distribution of the bet increase significantly differed between the simulated (*M* = -0.16, *SD* = 0.06) and the observed (*M* = -0.16, *SD* = 0.18; $D(89) = 0.33$, $p < .001$) data with a larger increase for the observed by about .001. Thus participants in Experiment S1 appear to have displayed a greater metacognitive blindness than anticipated by the simulation.

Likewise, we investigated whether the observed adjustment in HC responses in the prospective failure trials could be replicated under the null hypothesis of having no retrospective awareness. Compared to the average simulated HC responses at the pre-prospective failure baseline (*M* = 2.25, *SD* = 0.48), the HC responses at the simulated prospective failure trial were significantly lower by about .0002 (*M* = 2.25, *SD* = 0.48; $t(88) = 2.16$, $p = .034$, Cohen's $d = 0.23$). Thus, the simulation suggests that without any retrospective awareness, participants would have adjusted their HC responses to a small degree. However, a Kolmogorov-Smirnov test showed that the distribution of the HC decrease significantly differed between the simulated (*M* < 0.01, *SD* < 0.01) and the observed (*M* = 0.24, *SD* = 0.40; *D*(89) = 0.78, *p* < .001), meaning our findings of a "partial retrospective awareness" during prospective failures was a reliable trend.

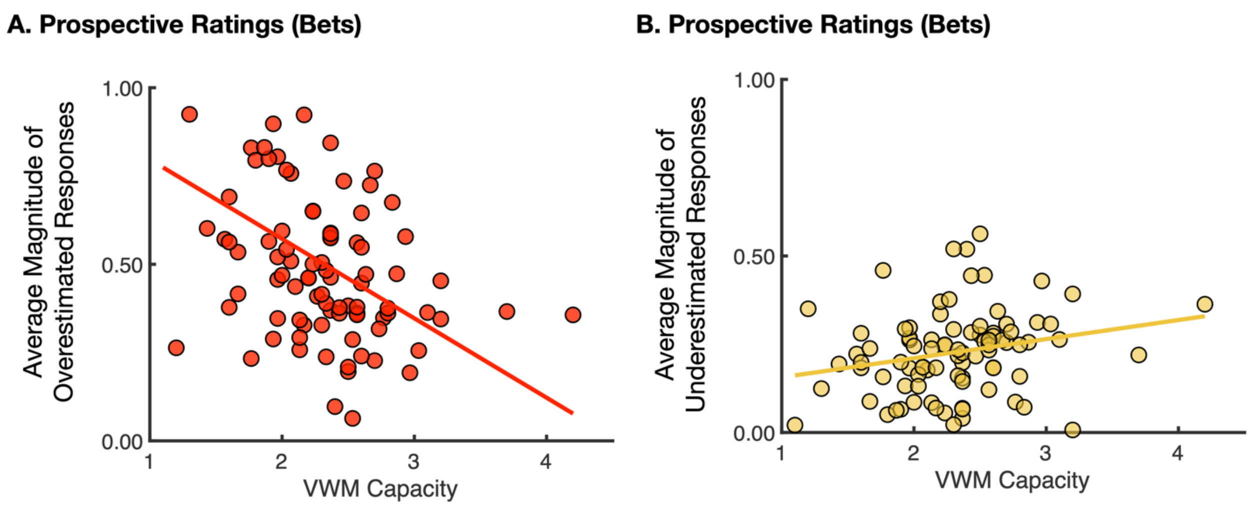
Simulation of post-prospective failure trial recall performance. After the initial prospective failure, we observed a sustained reduction in VWM performance. Thus, in this simulation, we also compared the recall performance for the simulated post-failure trials to that of pre-failure trials. In contrast to our observed results, the simulated post-failure trials showed miniscule but statistically significant improvement in recall performance for three consecutive trials (*M* = 2.41, 2. 41, 2. 41; *SD* = 0.32, 0.32, 0.33 for +1, +2, and +3 trials, respectively) compared to the simulated pre-failure baseline performance (*M* = 2.34, *SD* = 0.31; *t*(88) > 15.71, *p* < .001, Cohen's *d* > 1.66). Thus, our simulation suggests that the sustained reduction in VWM performance following a prospective failure was a reliable observation. Importantly, however, these results do not imply whether the observed sustained reduction was a strategy or natural product of the initial drop in performance. Because we did not replicate the observed reduction in performance, we did not extend the analysis for the bets and HC responses after the prospective failure trial since such results cannot be compared to the observed trend.

Individual differences in prospective and retrospective awareness. In Experiments 1 and 2, we found that low-capacity individuals were more likely to overestimate their upcoming performance than high-capacity individuals. Thus, we set to replicate our previous findings in the capacity-related individual differences for prospective awareness. There was a strong correlation between VWM capacity and the average magnitude of overestimated bets, such that low-capacity individuals were more likely to overestimate their upcoming VWM performance than high-capacity individuals at a greater rate and magnitude (*r*(87) = -.40, *p* < .001; see Figure S9A).

Interestingly, we also found a weak correlation between VWM capacity and the average magnitude of underestimated bets, such that high-capacity individuals were more likely to underestimate their upcoming VWM performance at a greater rate and magnitude than low-capacity individuals ($r(87) = .23$, $p < .05$; see Figure S9B). Altogether, we found that low-capacity individuals continued to overestimate their VWM performance at a greater average magnitude than high-capacity individuals in Experiment S1. However, introducing a greater range of to-be-remembered items may have influenced the metacognitive abilities of high-capacity individuals, such that they became more conservative in their bets.

Figure S9

Individual Differences in Over- and Under- Estimations in Experiment S1

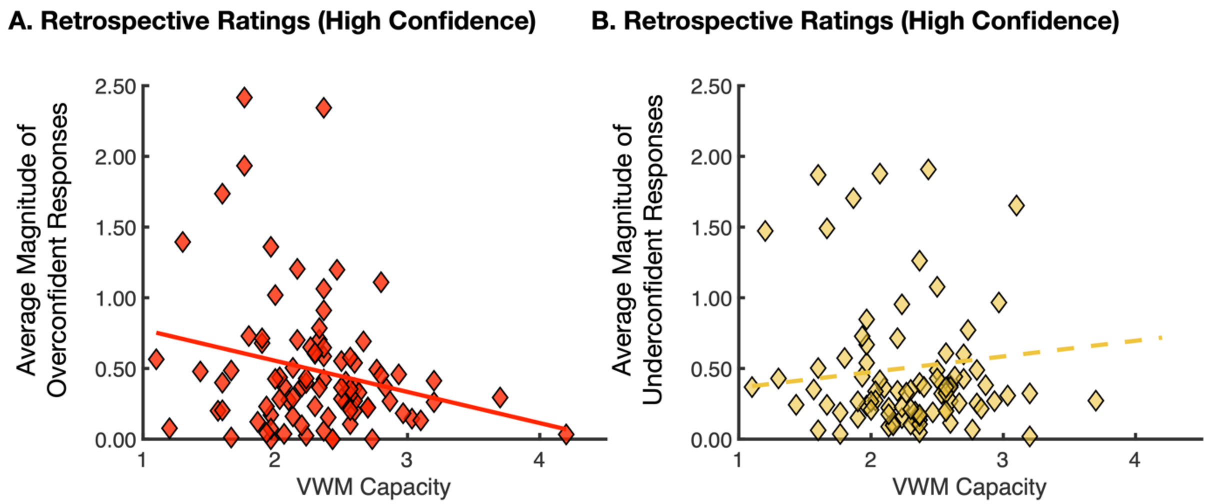


Note. $N = 89$. Significant findings denoted by solid lines. (A) The average magnitude of overestimated responses was calculated in two steps. First, we identified the bets participants made when they earned zero points, then subtracted the corresponding expected number of correctly reported from our Monte Carlo simulation. Second, we took the absolute sum of overestimations magnitudes for each individual and divided it by each individual's total number of trials. Together, this metric illustrates the average magnitude of the overestimation that individuals made on a given trial. (B) The average magnitude of underestimated responses was calculated with a similar method, but with trials in which participants' bets were less than the number of correctly reported items.

In addition to overestimations, we found that low-capacity individuals were more likely to be overconfident in their VWM performance than high-capacity individuals in Experiments 1 and 2. Thus, we next set to replicate our findings in the individual differences for reflections. Most critically, we found a correlation between VWM capacity and the average magnitude of overconfident trials, such that low-capacity individuals were more likely to be overconfident in their performance at a greater rate and magnitude than high-capacity individuals ($r(87) = -.23$, $p < .05$; see Figure S10A). Moreover, there was no correlation between VWM capacity and the average magnitude of underconfidence, such that low- and high- capacity individuals were equally likely to be underconfident in their VWM performance ($r(87) = .10$, $p = .328$; see Figure S10B). Altogether, we found that low-capacity individuals remained more overconfident in their VWM performance than high-capacity individuals in Experiment S1, but there was no influence of the larger set-size on capacity-related differences in underconfidence.

Figure S10

Individual Differences in Over- and Under- Confidence in Experiment S1

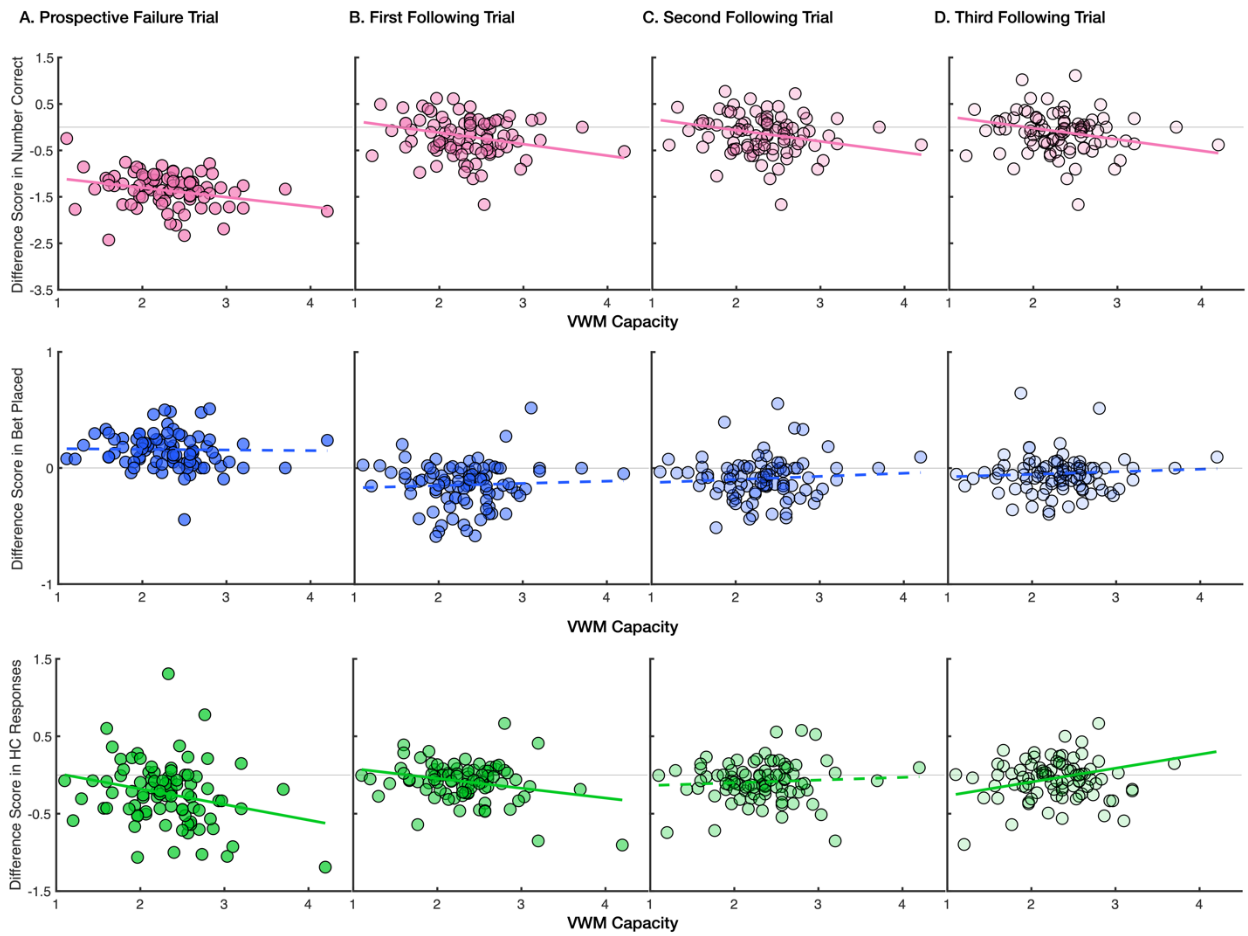


Note. *N* = 89. Significant findings denoted by solid lines. (A) The average magnitude of overconfident responses was calculated by the same method as our overestimation metric but using the total number of HC responses. (B) The average magnitude of underconfident responses was calculated with a similar method, but with trials in which participants' total HC responses were less than the number of correctly reported items.

Individual differences during and after prospective failures. In Experiments 1 and 2, we investigated capacity-related individual differences in performance and metacognitive awareness both at the moment of a prospective failure and after a prospective failure. There, we calculated a difference score between a participant's pre-prospective failure baseline responses and the trial of interest's responses. Thus, we similarly investigated whether there were any capacity-related differences in behaviors in Experiment S1 (see Figure S11). We found significant negative correlations between VWM capacity and the difference scores in the number of correctly recalled items during a prospective failure trial ($r(87) = -.23$, $p = .027$), first following trial ($r(87) = -.25$, $p = .016$), second following trial ($r(87) = -.25$, $p = .017$), and third following trial ($r(87) = -.26$, $p = .016$). These results suggest that high-capacity individuals suffered a greater reduction in performance during a prospective failure and were slower to recover their performance in the following trials compared to low-capacity individuals.

Figure S11

Experiment S1: Individual Differences at Prospective Failure and Post-Failure



Note. N = 89. Solid lines indicate a significant correlation. The difference scores were determined by subtracting each individual's pre-failure baseline response from the trial of interest's high response (e.g., VWM performance at a prospective failure trial– pre-failure baseline VWM performance). (A) Difference scores between the prospective failure trial and pre-failure baseline responses. (B) Difference score between the first following trial of a prospective failure and pre-failure baseline. (C) Difference score between the second following trial of a prospective failure and pre-failure baseline. (D) Difference score between the third following trial of a prospective failure and pre-failure baseline.

Next, we investigated whether there were any capacity-related differences in the bet and HC difference scores surrounding the prospective failure. We found no correlations between VWM capacity and the bet difference scores during a prospective failure trial ($r(87) = -.02$, $p = .885$), the first following trial ($r(87) = .05$, $p = .634$), second following trial ($r(87) = .08$, $p = .481$), nor the third following trial ($r(87) = .05$, $p = .622$). Thus, in this experiment, both high- and low- capacity individuals equally contributed to the bet increase at a prospective failure, and similarly adjusted their bets in the corresponding post-failure trials.

Lastly, we found a significant negative correlation between VWM capacity and the difference score of HC responses during a prospective failure trial ($r(87) = -.25$, $p = .017$), suggesting that high-capacity individuals reduced their HC responses at a greater magnitude than low-capacity individuals during the prospective failure. However, there was no significant correlation between VWM capacity and the total number of HC responses at the prospective failure trial ($r(87) = .06$, $p = .563$). There was a significant negative correlation at the first following trial ($r(87) = -.27$, $p = .012$), meaning high-capacity individuals continued to sustain the HC reduction more than low-capacity individuals, and that low-capacity individuals recovered to their pre-failure baseline HC responses faster than high-capacity individuals. While we did not find a significant correlation in the second following trial ($r(87) = .07$, $p = .508$), we found a significant correlation in the third following trial, $r(87) = .29$, $p = .007$. Together, our difference score analysis revealed that low-capacity individuals were more likely to revert to their pre-failure baseline HC responses than high-capacity individuals by the first following trial.

5 Individual differences in trial counts (Experiments 1, 2, and S1)

In this experiment, participants received performance-contingent points based on the accuracy of their bet to their VWM performance. Importantly, participants had to collect 600 points in each of four blocks to complete the experiment. The average number of trials taken to earn 600 points in each block, and each experiment, is shown in Table 1. Since these points somewhat corresponded to VWM performance, we investigated whether there were any VWM capacity-related differences in the number of trials taken to earn 600 points. Specifically, we determined the correlation between the VWM capacity estimates from the whole-report baseline task and the total number of trials taken to complete an experimental block (see Table 2). We found that low-capacity individuals had a greater number of trials in each block than high-capacity individuals across all three experiments.

Table 1

Average Trial Count Per Block

Experiment	Block 1	Block 2	Block 3	Block 4
1 (*N* = 87)	28.11 (7.10)	26.67 (7.04)	26.26 (7.00)	25.67 (7.65)
2 (*N* = 85)	31.25 (8.95)	30.25 (8.44)	28.75(6.79)	28.69 (7.39)
S1(*N* = 89)	29.48 (7.15)	27.33 (6.49)	26.83 (6.74)	26.35 (6.32)

Note. Mean trial count (*SD*) in each experimental block between the three experiments.

Table 2

Correlation Between VWM Capacity and Trial Counts

Block	Experiment 1	Experiment 2	Experiment S1
1	-.68***	-.59***	-.60***
2	-.61***	-.55***	-.60***
3	-.62***	-.48***	-.55***
4	-.66***	-.49***	-.54***

Note. Pearson correlation (r) between VWM capacity estimates and trial counts in each experimental block between the three experiments.
*** $p < .001$
** $p < .01$
* $p < .05$

6 Length of the sustained reduction in VWM performance (Experiments 1 and 2)

In both experiments, participants did not recover to their pre-prospective failure baseline VWM performance following a prospective failure. However, this sustained reduction in VWM performance appears longer in Experiment 1 (see Figure 5) than in Experiment 2 (see Figure 10). Thus, we aimed to investigate whether there was a statistical difference in this sustained reduction between experiments through a between-subjects repeated measures ANOVA (see Figure S12). First, we identified the difference score in VWM performance for each individual by subtracting their "pre-prospective failure VWM performance" from their VWM performance in each trial of interest. In this analysis, we included the VWM performance difference score for the prospective failure trial, the first following trial of a prospective failure, second following trial, and third following trial. We termed these variables as "Trial". Then, we defined a between-subjects factor of "Experiment" (presence of feedback versus absence of feedback).

We found a significant main effect of Trial ($F(3,510) = 1657.63$, $p < .001$, $\eta_p^2 = .91$), and a main effect of Experiment ($F(1,170) = 11.53$, $p < .001$, $\eta_p^2 = .06$). Post-hoc Tukey tests revealed that the main effects were driven by differences in the second and third following trials, but not at the prospective failure trial and first following trial. Specifically, at the moment of a prospective failure and its immediate following trial, participants in Experiments 1 and 2 suffered similar reductions in performance. However, at the second following trial, participants in Experiment 2 had a greater difference score (M = -0.26, SD = 0.47) than participants in Experiment 1 (M = -0.07, SD = 0.32; $t(170) = 3.87$, $p < .001$, Cohen's d = 0.59). Thus, in the absence of trial-by-trial feedback, VWM performance was poorer during the second following trial. This trend continued into the third following trial of a prospective failure, such that participants in Experiment 2 had a greater difference score (M = -0.25, SD = 0.44) than participants in Experiment 1 (M = -0.02, SD

= 0.33; $t(170) = 4.31$, $p < .001$, Cohen's $d = 0.66$). These results seem to suggest that the absence of feedback in Experiment 2 further delayed the recovery of VWM performance. However, given an insignificant interaction between Trial and Experiment, we encourage caution in interpreting these results ($F(3,510) = 1.91$, $p = .126$, $\eta_p^2 = .01$).

Figure S12.

Recall Performance Around Prospective Failures in Experiment 1 and Experiment 2