Divergence of Metacognition and Performance under Increasing Task Demand in Multiple-Object Tracking

Zoe Loh (zloh@ucmerced.edu)[1] & Spencer C. Castro[1]
Management of Complex Systems Department, University of California Merced

Abstract

As task demands increase, fewer cognitive resources remain for executing a task and monitoring one's own performance. This study investigates the point at which performance and metacognition (self-assessment) begin to diverge under increasing mental demand. Participants completed a dynamic visual tracking task in which they monitored target dots moving among distractors on a screen. Task difficulty was manipulated by varying the number of target dots from two to four, among a total of eight dots. Participants were prompted to estimate their performance during the trial, prior to receiving feedback. After each trial, they also completed the Paas scale to assess perceived cognitive workload. Participants accurately assessed the quality of their performance during the task and detected changes in workload afterward. They were less accurate in their judgements at higher difficulty levels. This suggests that as task demands increase, an individual's ability to monitor task performance is reduced.

Keywords: metacognitive monitoring; visual attention; working memory; multiple-object tracking; visual tracking

Background

Researchers in various disciplines study *metacognition*, the process of thinking about thinking (Jacobs & Paris, 1987). One form of metacognition is monitoring one's own ability to complete a task or goal (Flavell, 1979). Nelson and Narens (1990) propose that metacognitive processes occur between two levels: the *object* level, where task performance occurs, and the *meta* level, which represents and evaluates the performance of the object level. Metacognitive monitoring refers to the process of meta-level changes in response to actions at the object level, whereas control refers to adjustments to behavior based on that evaluation (Nelson & Narens, 1990). Here, we focus on monitoring as the process of evaluating performance during ongoing task execution and determining whether task demands exceed available cognitive resources.

One way to think about metacognitive monitoring would be to consider the processes of observing performance as a related but independent task. When there are multiple tasks at hand, how do we know what tasks to devote our resources to? In the *threaded cognition* model, there is a serial cognitive processor that coordinates task "threads" (Salvucci & Taatgen, 2008; Strayer et al., 2022). These task threads can be greedy or polite; a greedy thread will capture the cognitive resources required to complete the task quickly while a polite thread will wait for cognitive resources to become available (Salvucci & Taatgen, 2008). The threaded cognition model assumes that only one thread is being processed at a given time (Salvucci & Taatgen, 2008). Therefore, what people perceive as multi-tasking, or performing tasks simultaneously, is the result of rapid switching between task threads. In the *3CAPS* model, a pool of resources dedicated to executive functions is allocated to coordinate between tasks and check on performance level (Just et al., 2003). Dividing attention across multiple tasks increases task demands, which increases cognitive workload (Castro, 2017; Castro et al., 2019; Howard et al., 2020). As task demands increase, there are fewer resources to devote towards both performing the task and keeping track of task performance (Just et al., 2003). However, the change in workload due to increased demand from multi-tasking differs from increasing the difficulty of a single task (Howard et al., 2020). People exhibit greater response caution for multi-tasking but not increasing difficulty (Castro et al., 2023; Howard et al., 2020). If multi-tasking differs from increased task difficulty and metacognitive monitoring is inherent in any goal-driven task, this continuous performance assessment cannot be exactly like multi-tasking.

However, even if metacognitive monitoring differs from true multi-tasking, this does not mean that its processes remain independent from the resources utilized by the primary task. Previous studies have expanded metacognitive monitoring beyond its origins in learning and memory (Koriat, 2007) to these goal-driven cognitive processes of attention performance (Kessel et al., 2014) and working memory (Bryce et al., 2023). Researchers still debate the extent to which metacognitive and executive processes share resources, resulting in interference or performance reductions (Adam & Vogel, 2017; Bryce, et al., 2023). Other findings suggest that some dissociation allows for metacognitive judgements to remain stable despite declines in task performance under increased cognitive load in certain contexts (Conte, 2023). Therefore, in the present study, we expand on existing research by examining the impacts of a real-time performance assessment in a dynamic visual tracking task on attentional demands and a post-task cognitive workload measure.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

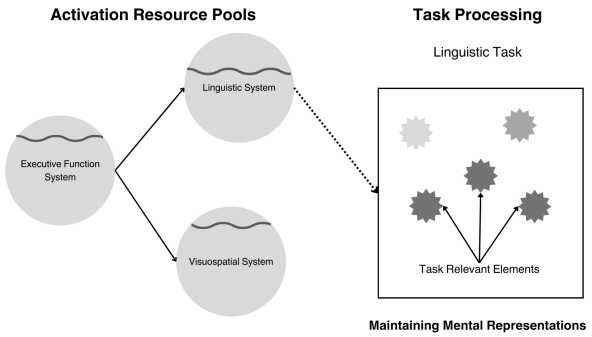

Figure 1: A visualization of the *3 Caps* model (Just et al., 2003). An executive function system allocates limited resources to the linguistic system for processing.

Multiple-Object Tracking

The multiple-object tracking (MOT) task was created as a test of ability to track multiple independent targets moving around the visual field. (Pylyshyn & Storm, 1988). This task requires participants to keep track of a subset of target objects among distractors throughout a trial. Completing an MOT task requires sustained visual attention and visual working memory (Fougnie & Marois, 2006). The objects to be tracked range in complexity; some versions use plus signs (Pylyshyn & Storm, 1988) or dots (Innes et al., 2021), while others use 3-dimensional sharks (Zelinsky & Neider, 2008) or cartoon animals (Drew et al., 2013). Tracking accuracy is impacted by capacity (number of targets), crowding (space between targets), hemifield placement (whether targets are in the same hemifield of the screen), and speed (how fast the targets are moving) (Scimeca & Franconeri, 2015). Previous studies employing an MOT task have examined the relationship between tracking accuracy and variables such as fluid intelligence (Tullo et al., 2018), motor control abilities (Wierzbicki et al., 2024), and response to a secondary task (Innes et al., 2021). Still, less is known about the link between behavioral outcomes of the task and online metacognitive processes.

Present Study

The present study aims to investigate the relationship between behavioral outcomes and self-assessment of performance for a dynamic tracking task under increasing workload demands. Here, we examine how metacognitive monitoring (Nelson & Narens, 1990), or the alignment between mental representations of performance and measured task response accuracy, is impacted when the number of targets increases for a multiple-object tracking task (Pylyshyn & Storm, 1988). Prior work suggests that increasing load may reduce response accuracy and, in some cases, metacognitive monitoring accuracy (Adam & Vogel, 2017; Bryce et al., 2023), particularly when both processes rely on shared executive resources (Conte, 2023). We hypothesize that behavioral outcomes and self-assessment of performance will decouple as the multiple-object tracking task becomes more difficult, making the self-assessments less predictive of successful target identification. We propose that this change reflects reduced metacognitive monitoring fidelity due to shared resources.

Method

Participants

Participants were recruited using the Prolific platform (Prolific Team, n.d.) and compensated for their time at a rate of $15 USD per hour. 79 participants were initially recruited to participate in the study. After excluding participants that failed the attention check or did not complete the study, 50 responses remained. On average, participants were 41 years old (SD = 11.64). 64% (n = 32) identified as male, 34% (n = 17) identified as female, and 2% (n = 1) identified as non-binary or third gender. In terms of education level, participants reported that they had completed or were currently pursuing: a bachelor's degree (50%, n = 25), an associate's degree (20%, n =10), a high school diploma (10%, n = 5), a master's degree (8%, n = 4), some college but no degree (8%, n = 4), a doctorate degree (2%, n =1), and trade school (2%, n =1). Two participants were excluded from further analysis due to data quality concerns because they had multiple trials time-out with no response.

Multiple-Object Tracking (MOT) Task

In this study, participants completed a multiple-object tracking (MOT) task adapted from Kim (2019). The purpose of an MOT task is to evaluate the ability to continually track multiple moving objects (Pylyshyn & Storm, 1988). During this task, participants tracked white dots moving on a gray screen among distractors. Task difficulty was manipulated by changing the number of dots that the participant is required to track each trial, ranging from 2 to 4 out of 8 total dots. At the start of each trial, the target dots flashed green (RGB 0, 128, 0) to distinguish them from the white (RGB 255, 255, 255) distractors. Then, the target dots stopped flashing to match the distractors, and all dots began moving around the screen. At the end of each trial, the dots stopped moving and participants were asked to click on the target dots. To discourage guessing, participants could not change their response after selecting a dot as a target. The number of allowed responses was equal to the number of targets for each trial. For instance, a trial with 3 targets only allowed participants to select 3 dots in their response.

Outcome Measures

Outcome measures included online perceived performance, post-task perceived cognitive workload, response accuracy, and response time. Perceived performance was measured using a 5-point Likert-type single item administered during task execution at a randomized time point (see Experiment Design and Procedure). Participants rated their performance on the current trial on a scale ranging from "very poor" to

"very good." Participants rated their perceived cognitive workload on the Paas scale (Paas, 1992) after every trial. The Paas scale is a 9-point single item scale that asks participants to report how much mental effort the task required ranging from "very, very low mental effort" to "very, very high mental effort" (Paas, 1992). As participants completed many trials, we employed the Paas scale as a simpler measure of cognitive workload compared to other self-report measures such as the NASA-TLX (Hart & Staveland, 1988) to reduce response fatigue. Response accuracy was calculated per trial as the proportion of correctly identified targets out of the total number of targets. First response time was calculated as the time elapsed between the start of the response screen and the first selected dot of the trial. Average response time per target was calculated by subtracting the time of the last dot selected from the time of the first dot selected and then dividing by the total number of responses in that trial.

Experiment Design and Procedure

The study was created and presented using the Gorilla Experiment Builder (www.gorilla.sc; Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed, 2018). Data was collected between September 24, 2025 and October 1, 2025. Participants completed three variations of the MOT task over the course of the experiment: pause report trials, pause control trials, and no pause trials. During pause-report trials, the moving dots temporarily froze and participants were prompted to indicate how well they believed they were performing using a single-item Likert-type scale presented above the display. The timing of the pause was randomized according to a Poisson distribution with a mean of 5000 ms to prevent participants from anticipating when it would occur. After participants responded, target dots were briefly re-cued in a different color before the trial resumed to minimize the pause's impact on performance. To obtain a baseline for the effect of the pause, some trials contained a pause with no question. These *pause control trials* were identical to the pause report trials, except participants were prompted to click a button to continue the trial for the pause. Finally, some trials were *no pause trials*. In these trials, the task proceeded without any interruption for the duration of the trial. These trials were included to ensure that participants were attending to the targets at the beginning of the trial rather than relying on the targets being re-cued after a pause occurred.

Participants were required to complete the study on a desktop computer. After providing consent to take part in the study, participants first answered demographic questions and read through instructions explaining the task. They then completed three guided practice trials, with one practice trial per type. After the practice trials block, participants completed a total of 50 trials for the main experiment. To minimize data loss, 30 (60%) of the trials were pause report trials, 15 (30%) were pause control trials, and 5 (10%) were no pause trials. Pause report and pause control trials were evenly split between the three difficulty levels (2, 3, and 4 targets). For the no pause condition, there were two 2-target trials, two 3-target, and one 4-target trial. Trial order was randomized to prevent potential order effects from difficulty or trial type.

Analysis

Analysis was conducted in R (R Core Team, 2021). Data from practice trials was excluded from the main analysis. Response accuracy scores were z-scored using the grand mean. Following the work of Bryce and colleagues (2023), cumulative link linear mixed effects models were used to analyze perceived performance and perceived cognitive workload as ordinal outcome measures. Z-scored response accuracy and difficulty level (i.e., 2 targets, 3 targets, or 4 targets) were set as fixed effects with by-participant intercepts to account for repeated measures. According to Berger and Kiefer, (2021), excluding reaction time outliers using cutoffs at the 5th and 95th percentiles introduces the least absolute bias. Therefore, we used this outlier identification method to exclude trials with first response reaction times below the 5th percentile (210.90 ms) and above the 95th percentile (2483.21 ms) from the reaction time data analysis.

Results

Response Accuracy Across Trial Types

Mean response accuracy was 0.74 (SD = 0.29) for no pause trials, 0.75 (SD = 0.26) for pause control trials, and 0.75 (SD = 0.27) for pause report trials. Model results revealed that participant response accuracy did not significantly differ between no pause trials and pause control trials (b = 0.03, SE = 0.07, t = 0.37, p = .71, 95% CI [-0.11, 0.16])) or between no pause trials and pause report trials (b = 0.03, SE = 0.06, t = 0.48, p = .63, 95% CI [-0.10, 0.16]).

Response Accuracy Across Difficulty Levels

Mean response accuracy was 0.90 (SD = 0.26) for 2 targets, 0.73 (SD = 0.27) for 3 targets, and 0.63 (SD = 0.21) for 4 targets. Model results revealed that participant response accuracy decreased as the number of targets increased from 2 targets to 3 targets (b = -0.61, SE = 0.05, t = -11.45, p < .001, 95% CI [-0.72, -0.51]) and from 2 targets to 4 targets (b = -0.98, SE = 0.05, t = -18.34, p < .001, 95% CI [-1.08, -0.87]).

Response Time Across Difficulty Levels

Mean reaction time for the first response of each trial was 748.62 ms (SD = 396.36) for 2 targets, 798.31 ms (SD = 386.82) for 3 targets, and 867.42 ms (SD = 416.09) for 4 targets. The reaction time for the first response of each trial increased as task difficulty increased to 3

targets (b = 0.07, SE = 0.03, t = 2.57, p = .01, 95% CI [0.02, 0.13]) and to 4 targets (b = 0.15, SE = 0.03, t = 5.50, p < .001, 95% CI [0.10, 0.21]). Mean reaction time per response in a trial was 475.68 ms (SD = 183.82) for 2 targets, 606.95 ms (SD = 244.27) for 3 targets, 684.71 ms (SD = 256.85) for 4 targets. Average reaction time per response also increased as difficulty increased to 3 targets (b = 0.24, SE = 0.02, t = 14.03, p < .001, 95% CI [0.21, 0.28]) and 4 targets (b = 0.37, SE = 0.02, t = 21.40, p < .001, 95% CI [0.34, 0.40]).

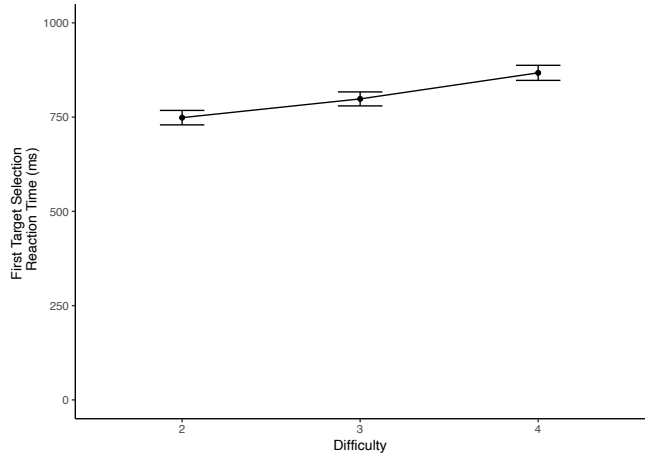

Figure 2: The response time in milliseconds of selecting the first target to 2, 3, and 4 targets respectively. Error bars are 95% confidence intervals around the mean utilizing the Cousineau-Morey method (Cousineau, 2005; Morey, 2008).

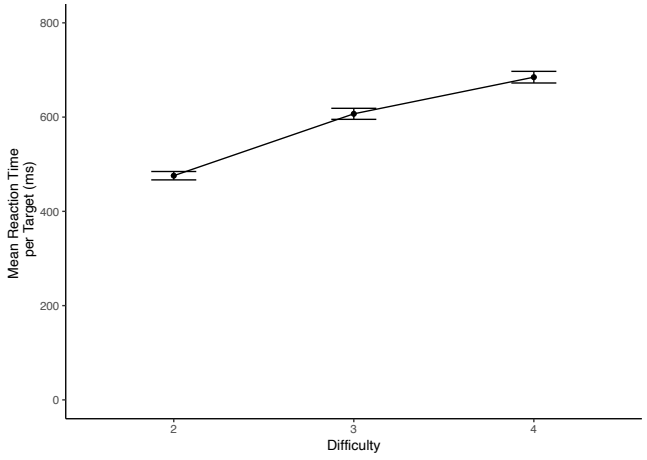

Figure 3: The response time in milliseconds to 2, 3, and 4 targets respectively averaged by trial. Error bars are 95% confidence intervals around the mean utilizing the Cousineau-Morey method (Cousineau, 2005; Morey, 2008).

Perceived Performance and Response Accuracy

Mean perceived performance score was 3.74 (SD = 0.92) for 2 targets, 3.31 (SD = 0.92) for 3 targets, and 2.70 (SD = 1.04) for 4 targets. Model results revealed a main effect of difficulty on perceived performance. As the difficulty of the task increased from 2 targets to 3 targets, participants were less likely to rate themselves as performing well on the task (b = -0.88, SE = 0.14, z = -6.19, p < .001, 95% CI [-1.16, -0.60]). Similarly, there was a decrease in perceived performance when task difficulty increased from 2 targets to 4 targets (b = -2.37, SE = 0.16, z = -14.78, p < .001, 95% CI [-2.68, -2.06]). Response accuracy was a statistically significant predictor of perceived performance (b = 0.58, SE = 0.11, z = 5.50, p < .001, 95% CI [0.38, 0.79]). For every standard deviation unit increase in response accuracy, participants were 1.79 times as likely to rate higher on the perceived performance scale. The relationship between response accuracy and perceived performance changed as the difficulty increased to 3 targets (b = -0.44, SE = 0.14, z = -3.29, p = .001, 95% CI [-0.71, -0.18]) and 4 targets (b = -0.75, SE = 0.15, z = -4.91, p < .001, 95% CI [-1.05, -0.45]). Participant ratings of their performance became more uncertain as the difficulty level increased.

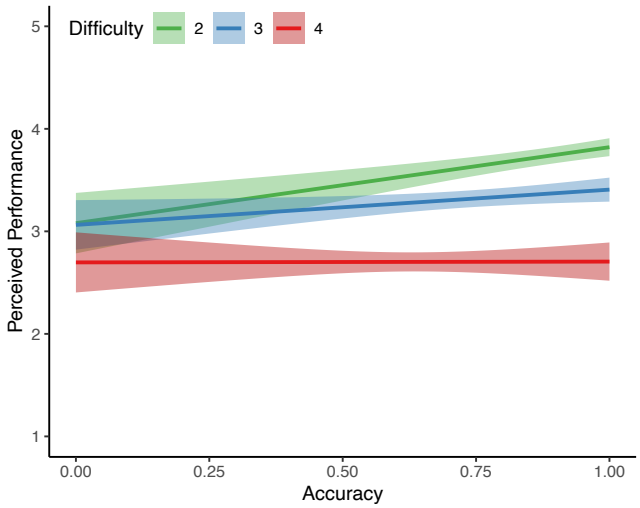

Figure 4: The average rating of performance on the 5-point perceived performance scale by the accuracy of the trial responses. The shaded areas are 95% confidence intervals for predicted values utilizing the Locally Estimated Scatterplot Smoothing (LOESS) method (Wickham, 2016).

Perceived Workload and Response Accuracy

The mean perceived workload rating was 5.57 (SD = 2.11) for 2 targets, 6.94 (SD = 1.47) for 3 targets, and 7.64 (SD = 1.42) for 4 targets. Model results revealed a main effect of difficulty on perceived workload. Participants rated the task as requiring greater mental effort as the difficulty level increased from 2 targets to 3 targets (b = 1.61, SE = 0.14, z = 11.58, p < .001, 95% CI [1.34, 1.88]) and 4 targets (b = 3.20, SE = 0.17, z = 19.10, p < .001, 95% CI [2.87, 3.52]). Response accuracy was also a statistically significant predictor of perceived workload (b = -0.68, SE = 0.10, z = -6.63, p < .001, 95% CI [-0.89, -0.48]). For every

standard deviation unit increase in response accuracy, participants were 0.5 times as likely to rate higher on the Paas scale; in other words, as response accuracy increased, perceived workload decreased. The relationship between response accuracy and perceived workload changed as the difficulty increased to 3 targets ($b = 0.39$, $SE = 0.13$, $z = 3.06$, $p = .002$, 95% CI [0.14, 0.63]) and 4 targets ($b = 0.61$, $SE = 0.16$, $z = 3.86$, $p < .001$, 95% CI [0.30, 0.91]). Workload ratings became more uncertain as task demands increased.

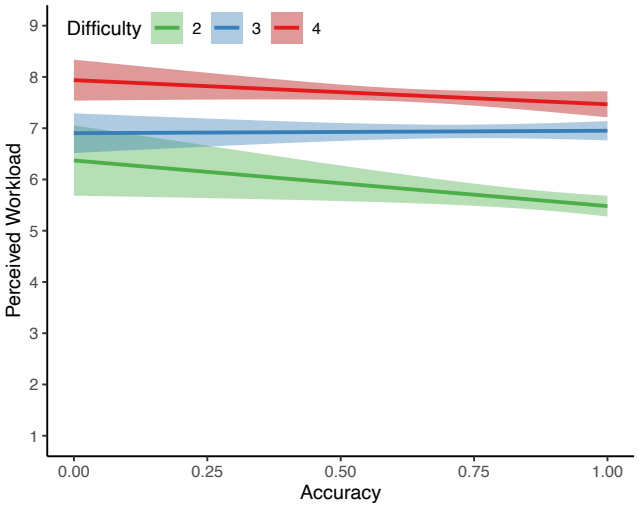

Figure 5: The average rating of cognitive workload on the Paas scale (Paas, 1992) by the accuracy of the trial. The shaded areas are 95% confidence intervals for predicted values utilizing the Locally Estimated Scatterplot Smoothing (LOESS) method (Wickham, 2016).

Discussion

In this study, we examined how increasing task demands can influence the relationship between behavioral outcomes and self-assessment of performance. We found that participants could assess the quality of their responses and detect changes in workload overall. However, the strength of the relationship between response accuracy and perceived performance decreased as the difficulty of the task increased. Our work revealed a similar decoupling between response accuracy and perceived workload between difficulty levels. Participant response time increased as the task became more difficult, both for the first selected response of a trial and for each target on average. Overall, these findings aligned with our prediction that the relationship between perceived performance and behavioral outcomes would shift, reflecting reduced individual metacognitive monitoring ability, as the task becomes more cognitively demanding.

Metacognitive Monitoring Accuracy

The outcomes of our study extend the findings reported in Bryce and colleagues (2023): individuals maintained high monitoring accuracy for their responses, but increases in working memory demands decrease monitoring accuracy. We further demonstrated that metacognitive monitoring–performance decoupling emerges more strongly in continuous tracking tasks where monitoring must be maintained online. One potential explanation for this pattern is that metacognitive monitoring and the cognitive resources used to complete the MOT task (i.e., visual attention and working memory) drain a shared resource pool (Just et al., 2003). This view is similar to those discussed by Bryce and colleagues (2023) and Adam and Vogel (2017). In contrast, Conte and colleagues (2023) did not identify a difference in metacognitive accuracy between memory load conditions during an n-back task. They suggest that there may be a dissociation between the mechanisms for working memory and metacognitive judgement. The current finding places constraints on a dissociation account (e.g., Conte et al., 2023), which would predict more stable monitoring under conditions where memory representations are not discretely encoded. It is possible that the difference in results is specific to aspects of the cognitive task being evaluated.

At the level of difficulty, some strategy differences may have played a role in determining accuracy distributions. For example, the number of total targets remained constant at eight, meaning that the probability of guessing a target went from one in four (i.e., 25%) to one in two (i.e., 50%) over the difficulty levels. Participants rarely missed all of the targets in the 4-target condition despite having a lower average accuracy. They could have either selected a subset of the four targets to increase their chances of successfully tracking the remaining targets, or their likelihood of selecting at least one target just increased over the difficulty levels. In future studies, we will have participants report strategies with a larger sample size to detect individual differences. Additionally, we will test individual differences in working memory with an operation span task.

Response Time

In addition to self-report measures, we also performed an exploratory analysis of task difficulty on response time. Interestingly, both first target selection response time and average response time per target selection increased with greater task demands. If all target positions in a trial are retrieved simultaneously from working memory at the start of responding, we would expect response time for the first response to increase with more targets, but response time per target to stay constant. Conversely, if the position of each target was retrieved serially, we would expect first response time to be similar across different numbers of targets. Here, our results do not exactly align with either of those possibilities.

Limitations

One limitation of the present study is we did not assess participant strategies in completing the task. Another limitation is that we did not include individual difference measures for participant characteristics such as working memory span and multi-tasking ability. We plan to address these limitations in a future study by asking participants to report their strategies and including assessments for individual differences.

Conclusion

In summary, by systematically manipulating task difficulty through increases in the number of target dots, this study demonstrates that individuals are generally capable of accurately monitoring their performance and perceived cognitive workload. Participants' pre-feedback performance estimates closely aligned with objective outcomes, and their Paas scale ratings reliably reflected changes in task demands. However, this monitoring ability deteriorated as task difficulty increased, with participants showing not only reduced accuracy in their judgments under higher cognitive load, but no relationship between their ratings and accuracy on a trial-to-trial basis. These findings suggest that although metacognitive awareness remains robust at lower levels of demand, increasing task complexity constrains individuals' capacity to effectively evaluate their own performance. These results demonstrate that the critical bottlenecks in attentional and working memory resources at least partially apply to self-monitoring under elevated workload. This study has potential implications for task demands when designing systems or environments that rely on users' accurate self-assessment.

Future studies could determine the mechanisms by which assessment during the task differs from assessment after the task, as well as how perceived *mental effort* differs from perceived performance.

Acknowledgments

Thank you to the members of the TECH lab for their support on this project. This work was funded by the UC President's Dissertation Year Fellowship.

References

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

Berger, A., & Kiefer, M. (2021). Comparison of Different Response Time Outlier Exclusion Methods: A Simulation Study. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.675558

Bryce, D., Kattner, F., Birngruber, T., & Wellingerhof, P. (2023). Monitoring accuracy suffers when working memory demands increase: Evidence of a dependent relationship. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(12), 1909–1922. https://doi.org/10.1037/xlm0001262

Castro, S. (2017). How Handheld Mobile Device Size and Hand Location May Affect Divided Attention. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1370–1374. https://doi.org/10.1177/1541931213601826

Castro, S. C., Heathcote, A., Cooper, J. M., & Strayer, D. L. (2023). Dynamic workload measurement and modeling: Driving and conversing. *Journal of Experimental Psychology: Applied*, *29*(3), 645–653. https://doi.org/10.1037/xap0000431

Castro, S. C., Strayer, D. L., Matzke, D., & Heathcote, A. (2019). Cognitive workload measurement and modeling under divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(6), 826–839. https://doi.org/10.1037/xhp0000638

Conte, N., Fairfield, B., Padulo, C., & Pelegrina, S. (2023). Metacognition in working memory: Confidence judgments during an *n*-back task. *Consciousness and Cognition*, *111*, 103522. https://doi.org/10.1016/j.concog.2023.103522

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*. https://doi.org/10.20982/tqmp.01.1.p042

Drew, T., Horowitz, T. S., & Vogel, E. K. (2013). Swapping or dropping? Electrophysiological measures of difficulty during multiple object tracking. *Cognition*, *126*(2), 213–223. https://doi.org/10.1016/j.cognition.2012.10.003

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Fougnie, D., & Marois, R. (2006). Distinct Capacity Limits for Attention and Working Memory: Evidence From Attentive Tracking and Visual Working Memory Paradigms. *Psychological Science*, *17*(6), 526–534. https://doi.org/10.1111/j.1467-9280.2006.01739.x

Howard, Z. L., Evans, N. J., Innes, R. J., Brown, S. D., & Eidels, A. (2020). How is multi-tasking different from increased difficulty? *Psychonomic Bulletin & Review*, *27*(5), 937–951. https://doi.org/10.3758/s13423-020-01741-8

Innes, R. J., Evans, N. J., Howard, Z. L., Eidels, A., & Brown, S. D. (2021). A Broader Application of the Detection Response Task to Cognitive Tasks and Online Environments. *Human Factors*, *63*(5), 896–909. https://doi.org/10.1177/0018720820936800

Jacobs, J. E., & Paris, S. G. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist*, *22*(3–4), 255–278.

Just, M. A., Carpenter, P. A., & Miyake, A. (2003). Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related

potential studies of brain work. *Theoretical Issues in Ergonomics Science*, *4*(1–2), 56–88. https://doi.org/10.1080/14639220210159735

Kessel, R., Gecht, J., Forkmann, T., Drueke, B., Gauggel, S., & Mainz, V. (2014). Metacognitive monitoring of attention performance and its influencing factors. *Psychological Research*, *78*(4), 597–607. https://doi.org/10.1007/s00426-013-0511-y

Kim, S. H. (2019). *Multiple Object Tracking Paradigm* [Computer software]. https://github.com/san-ha-kim/multiple-object-tracking-paradigm

Koriat, A. (2007). Metacognition and consciousness. In *The Cambridge handbook of consciousness* (pp. 289–325). Cambridge University Press. https://doi.org/10.1017/CBO9780511816789.012

Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Nelson, T. O., & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. In *Psychology of Learning and Motivation* (Vol. 26, pp. 125–173). Elsevier. https://doi.org/10.1016/S0079-7421(08)60053-5

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Prolific Team. (n.d.). *Prolific*. Retrieved October 20, 2025, from https://www.prolific.com

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spat Vis*, *3*(3), 179–197. https://doi.org/10.1163/156856888x00122

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, *115*(1), 101–130. https://doi.org/10.1037/0033-295X.115.1.101

Scimeca, J. M., & Franconeri, S. L. (2015). Selecting and tracking multiple objects. *WIREs Cognitive Science*, *6*(2), 109–118. https://doi.org/10.1002/wcs.1328

Strayer, D. L., Castro, S. C., & McDonnell, A. S. (2022). The Multitasking Motorist. In A. Kiesel, L. Johannsen, I. Koch, & H. Müller (Eds.), *Handbook of Human Multitasking* (pp. 399–430). Springer International Publishing. https://doi.org/10.1007/978-3-031-04760-2_10

Tullo, D., Faubert, J., & Bertone, A. (2018). The characterization of attention resource capacity and its relationship with fluid reasoning intelligence: A multiple object tracking study. *Intelligence*, *69*, 158–168. https://doi.org/10.1016/j.intell.2018.06.001

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wierzbicki, M., Rupaszewski, K., & Styrkowiec, P. (2024). Comparing highly trained handball players' and non-athletes' performance in a multi-object tracking task. *The Journal of General Psychology*, *151*(2), 173–185. https://doi.org/10.1080/00221309.2023.2241950

Zelinsky, G. J., & Neider, M. B. (2008). An eye movement analysis of multiple object tracking in a realistic environment. *Visual Cognition*, *16*(5), 553–566. https://doi.org/10.1080/13506280802000752