Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-021-01961-6

BRIEF REPORT

Using Past and Present Indicators of Human Workload to Explain Variance in Human Performance

Zachary L. Howard[1] · Reilly Innes[2] · Ami Eidels[2] · Shayne Loft[1]

Accepted: 20 May 2021

Abstract
Cognitive workload is assumed to influence performance due to resource competition. However, there is a lack of evidence for a direct relationship between changes in workload *within an individual* over time and changes in that individual's performance. We collected performance data using a multiple object-tracking task in which we measured workload objectively in real-time using a modified detection response task. Using a multi-level Bayesian model controlling for task difficulty and past performance, we found strong evidence that workload both during and preceding a tracking trial was predictive of performance, such that higher workload led to poorer performance. These negative workload-performance relationships were remarkably consistent across individuals. Importantly, we demonstrate that fluctuations in workload *independent from the task demands* accounted for significant performance variation. The outcomes have implications for designing real-time adaptive systems to proactively mitigate human performance decrements, but also highlight the pervasive influence of cognitive workload more generally.

Keywords Cognitive and attentional control · Attention and executive control · Bayesian statistics · Dual-task performance

Introduction

Understanding the causes of variance in human performance is a key goal of psychology. In particular, there is an urgent need to predict human performance decrements in a variety of industrial, transportation, military, and medical contexts, to design resilient work systems (Boehm-Davis, Durso, & Lee, 2015). If we can predict poor performance before it occurs, we may be able to intervene to increase workplace efficiency and prevent human error. Adaptive human-machine (automation) systems have been proposed to track information about task demands and operator state to modify work systems to proactively mitigate human performance deficits (Feigh, Dorneich, & Hayes, 2012). Although this concept has a long history (Rouse, 1988), there are challenges to practical implementation (Sauer, Kao, & Wastell, 2012). We propose that objectively measuring operator cognitive workload might better allow prediction of future operator performance.

The term *workload* describes the relationship between task demands and available human mental capacity, and is a critical construct for understanding performance (Young, Brookhuis, Wickens, & Hancock, 2015). High mental workload has been associated with accident risk in various work domains (e.g., Chen, Song, & Lin, 2016; Habib, Shalkamy, & El-Basyouny, 2019). There are a variety of subjective and objective measures of workload (Charles & Nixon, 2019; Matthews, Reinerman-Jones, Barber, & Abich, 2015; Thorpe, Nesbitt, & Eidels, 2020). In general, they are all based on the assumption that workload will vary as a function of task demand and the limited human information-processing resources available to meet those demands (Gopher, & Donchin, 1986). The relationship between workload and performance is complex, and performance does not always suffer with increased workload (Hancock & Matthews, 2019). This suggests that workload and performance are different constructs, and can dissociate (Wickens, 2008). However, most theories of workload predict performance impairments with increased workload, due to competition for shared resources (Young et al., 2015). Many prior studies linking workload and performance have used cross-sectional designs that aggregate repeated measurements from individuals to evaluate between-

✉ Zachary L. Howard
zach.howard@uwa.edu.au

[1] Department of Psychology, The University of Western Australia, Perth, WA, Australia

[2] Department of Psychology, University of Newcastle, Newcastle, NSW, Australia

Published online: 22 June 2021

person or between-task effects. While this level of analysis is often sufficient (e.g., assessing impact of training), it cannot indicate what level of performance to expect for a specific individual at a given workload. To our knowledge, a *real-time association* between objectively measured workload and performance has yet to be demonstrated, despite proponents of workload-based adaptive automation assuming this relationship exists.

We aimed to examine the extent to which an objective measure of workload could predict variation in individual performance. The primary task was a multiple-object-tracking task (Innes, Evans, Howard, Eidels & Brown, 2019; Pylyshyn & Storm, 1988), which is broadly representative of a range of work contexts such as air traffic control, train control, and maritime surveillance (Loft, Neal, Sanderson, & Mooij, 2007). Workload was objectively measured using a secondary task that required individuals to respond to transiently presented signals. We demonstrate that changes in an individual's workload both preceding and during the performance period of the tracking task can predict unique variance in performance after controlling for task difficulty and past performance.

Workload and performance

A fundamental tenant of cognitive psychology is the limited-capacity nature of human information processing (Kahneman, 1973; Norman & Bobrow, 1975). It is assumed that humans have a finite pool of cognitive resources to allocate to tasks (often termed *resource capacity*). The amount of resources (relative to capacity) needed to handle task load is known as cognitive or mental "workload." More demanding tasks require a higher level of "workload" on average (Young et al., 2015). Workload at the between-person level represents differences between individuals in resource capacity and the ability to self-regulate that capacity as a function of task demands, and also reflects individual differences in skill (Hockey, 1997). Higher average workload can indicate that, compared to others, an individual's capacity is exceeded, which can degrade performance compared to others.

Here, we are concerned with how fluctuations in an individual's workload relate to changes in their own moment-to-moment performance. There are many reasons an individual's workload may fluctuate independent of their task. Some theories posit that "mental resources" are a literal resource(s) that can be dynamically allocated (Kahneman, 1973; Wickens, 2002), and others extend this concept to assume these resources deplete over time (e.g., Baumeister, 2014; Hagger et al., 2010). Kurzban, Duckworth, Kable, and Myers (2013) suggest that the allocation of resources leads to an opportunity-cost to which the operator may be sensitive. It is also true that humans are prone to mind-wandering (Thompson, Besner, & Smilek, 2015), which reduces the attentional resources available. Other theories propose time-dependent workload increases (e.g., the well-known vigilance decrement has been attributed to workload decrements over time; Grier et al., 2003). Importantly, while theories may differ on the proposed *mechanisms* underlying fluctuating workload, almost all predict that workload will fluctuate, and that higher workload should lead to performance decrements. We consider whether and how performance changes when workload measurably fluctuates.

Surprisingly, only two studies have examined the relationship between workload and performance *in real-time* (as opposed to aggregated at a manipulation level, such as easy vs. difficult; Loft et al., 2018; Mracek, Arsenault, Day, Hardy, & Terry, 2014). Both showed that increased workload for an individual (i.e., increased workload relative to one's own average workload) was associated with a decrease in subsequent performance (i.e., decreased performance relative to one's average performance) in command and control tasks. However, these studies have several limitations, particularly that they assessed workload subjectively. Subjective workload ratings reflect the relationship between perceived task demands and a self-appraisal of resource capacity, and are subject to biases (e.g., social desirability) and self-awareness (Annet, 2002). Subjective measures can also be intrusive, and it can be difficult to obtain reliable estimates in real-time (which is a necessity for adaptive systems).

We examine whether a secondary task-performance measure of workload can predict moment-to-moment variation in within-person primary-task performance. The Detection Response Task (DRT; Innes et al., 2019; Young, Hsieh, & Seaman, 2013) provides an objective, continuous measure of residual human capacity that is assumed to be inversely related to the proportion of resources allocated to the primary task. DRT has been used successfully in many applied contexts, including driving (Strayer et al., 2015) and helicopter flight-simulators (Innes, Howard, Thorpe, Eidels, & Brown, 2020), to provide an objective measure of workload that taps into residual processing capacity (Palada, Neal, Strayer, Ballard, & Heathcote, 2019). In addition to being an objective measure of workload, the DRT is an analogue to ancillary activities associated with monitoring displays in work contexts (e.g., detecting event onsets). The DRT can thus tap into *real-time* fluctuations in workload, and can be naturally applied in many real-world tasks, making it an ideal choice for the present study.

We consider both workload concurrent with performance (reflecting current spare resource capacity) and workload *preceding* a given trial. Recent work suggests there can be lingering effects of workload for some time after completing a demanding task (Bowden, Loft, Wilson, J. Howard, & Visser, 2019), which might therefore be predictive of performance independent of direct resource conflict. This might

reflect depleted resources that take time to recover (e.g., Baumeister, 2014), ongoing sensitivity to opportunity costs (Kurzban et al., 2013), or slow reallocation of resources. We also include a between-subjects workload variable to determine whether people with a higher average workload perform worse overall (as suggested by, e.g., Hockey, 1997; Humphreys & Revelle, 1984). Further, in some cases a workload tool such as the DRT may simply not be available, and instead performance is interpreted as a workload proxy (Wickens, 2002). In many other cases, actual performance will not be known in real-time. We therefore include a prior-performance measure to allow assessment of how both workload *and* prior performance are related to current performance when both are known. This is an important contribution over and above prior work.

Our study presents several other important novel contributions. First, we control for performance decrements related to changes in task difficulty. This allows interpretations of the workload-to-performance relationship to be independent of task-related effects. Second, we allow individual-level slopes on our regression parameters. In this way, we can not only determine the "group," or "average," effect of individual workload on individual performance, but also determine the homogeneity of these individual-level effects (in other words, to determine whether or not there are individual differences in the effect of workload on performance). Conceptually, adaptive systems assume relative homogeneity in how humans are affected by workload. Individual differences would reduce the generalizability of adaptive systems, as they would need to be tailored to individual operators (beyond simply adjusting for mean performance), for example, if one operator performs better under load and another performs worse. If there are individual differences in the relationship between workload and performance these could be shown by variability in individual coefficients. Therefore, we provide a novel empirical test of the homogeneity of workload-performance relationships.

Method

Participants

A total of 133 undergraduate psychology students from the University of Newcastle participated in an online experimental session and were reimbursed with course credit. The sample was a convenience sample and total sample size was based on sign-ups within the study period (and is comparable or larger than similar previous studies, e.g., Howard et al., 2020). In order to minimize the potential for the identification of individual participants within our data set (openly available on the OSF at https://osf.io/6me5z/), our online data collection system was anonymous, and did not involve the collection of any potentially identifying demographic information. The study was approved by the Human Research Ethics Committee of the University of Newcastle.

Tasks

Participants completed a dual-task paradigm. The primary task was to track objects moving on the screen (MOT), with a secondary detection task appearing at the top-center of the experiment window at random intervals (DRT), as illustrated in Fig. 1. The timelines of the MOT and DRT are illustrated in Fig. 2. Both tasks were administered on the participant's computer concurrently, with the DRT stimulus displayed at the top of the MOT display area.

Moving object tracking (MOT) task

The MOT required participants to track the movement of zero, one, or four small target discs among a set of distractor dots. The number of disks to be tracked (tracking load) was considered a workload manipulation, validated in previous studies (Innes et al., 2019; Innes & Kuhne, 2020). There were always ten moving discs on display, and non-target (distractor) discs were colored red for the entire MOT trial. The to-be-tracked discs were initially colored blue to identify them as "targets." All dots moved around the display area for 15 s and did not require input from the participant. After 3 s from movement onset, the target dots changed color from blue to red, and the participants were required to track their motion. Each disc was

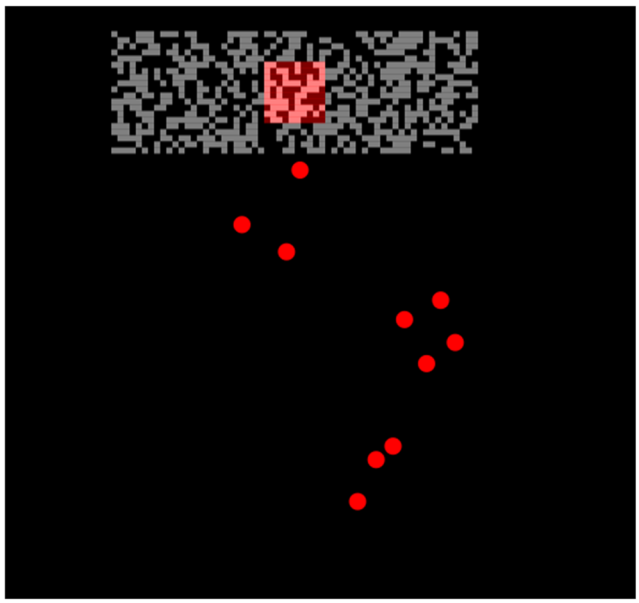

Fig. 1 Example of concurrent detection response task (DRT; red square) and moving object tracking (MOT; red dots) tasks. The red dots move around the display for a period of 15 s, while the DRT displays at 3- to 5-s intervals. The DRT requires an affirmative detection response while the MOT requires passive tracking during the 15-s period

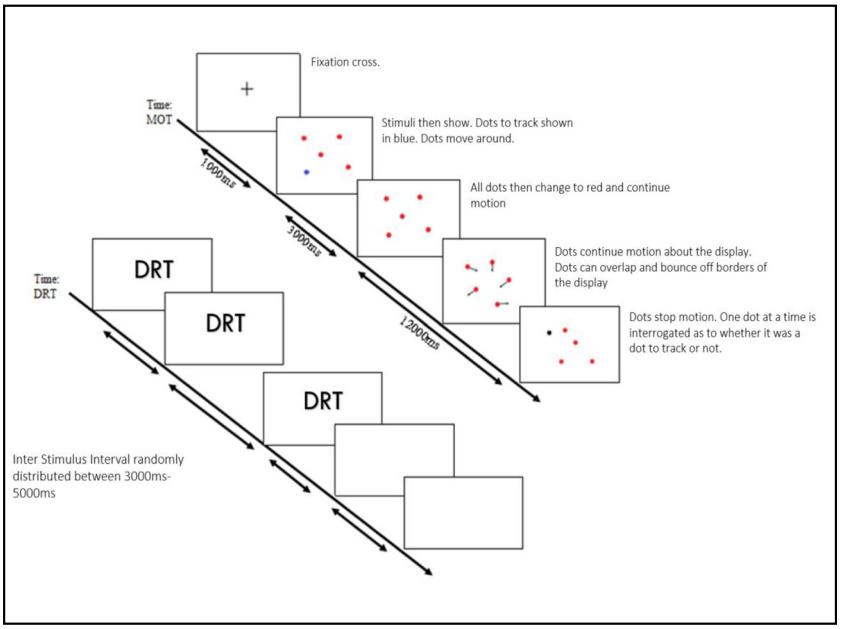

Fig. 2 Timeline of both concurrent detection response task (DRT) and moving object tracking (MOT) for all studies. Note the DRT stops at the MOT interrogation phase to ensure there is no response competition

circular with a diameter of 14 pixels, corresponding to a visual angle of approximately 2° at 60 cm viewing distance on a standard 24-in. monitor. Discs moved randomly within the display area and could overlap briefly if their paths crossed. If a disc reached the edge of the 150 x 150 pixel display area it bounced away randomly. An interrogation stage commenced following the 15 s, in which each of the ten dots, in turn, were colored white and participants were asked "was the white dot one of the targets?", with a yes/no response indicated using the keyboard ("O" and "P," respectively). These questions were self-paced and continued until all ten dots had been interrogated.

Detection response task (DRT)

To measure workload, we implemented a variation of the DRT developed in Innes et al. (2019). Our version of the DRT is administered according to ISO 17488 (Young, Hsieh, & Seaman, 2013) presentation intervals, but adapted to work within a JavaScript program. The DRT stimulus was a red square that appeared randomly in a rectangle above the MOT display (Fig. 1). The rectangle was embedded in noise consisting of uniformly sampled grey or black pixels, consistent with previous applications (Howard et al., 2020). Throughout the 15 s, the noise above the display was resampled 15 times per second. The red DRT square was displayed every 3–5 s (sampled uniformly). Participants were required to respond each time the DRT stimulus was displayed using the "T" key on the keyboard. The DRT stimulus remained on screen for 1 s unless the participant responded faster, and the maximum time to respond to the DRT stimulus was 2.5 s (button presses recorded after this time were considered misses), meaning participants could respond after the stimulus had disappeared, but there could be no overlap with the subsequent trial. The DRT display was not shown outside of the 15-s tracking periods, so there were no DRT presentations during instructions, breaks, or MOT "interrogation" phase.

Procedure

Participants were recruited online, and participated using their own device. Eighty participants were presented three tracking load levels – zero, one, and four dots. A further 53 participants were only presented with the one- and four-load levels. The zero-tracking condition was not of interest in the current study and was initially included for use as a "baseline" condition for other purposes (e.g., see Innes et al., 2019). As all other aspects of the designs were identical between the two task variants, we combined the data for the purpose of the following analyses. Participants first completed a practice block of five MOT trials, in which they were required to track two dots and simultaneously complete the DRT task. Following the practice, participants who were presented with three load levels were presented a total of nine blocks of 12 MOT trials (three blocks for each load level) for a total of 108 trials, each consisting of 15 s of tracking and the subsequent interrogation phase. The final 53 participants instead were presented ten blocks of 12 MOT trials (five blocks per load level) for a total of 120 MOT trials. In all cases the initial order of the load levels was randomized, and this ordering was repeated throughout the trial (this ensured that the tracking load

changed every block). Participants were instructed at the beginning of each block how many target dots were to be tracked for that block. As the onset time of the DRT presentation was random, the trial numbers slightly varied, but on average there were between three and six DRT trials in a 15-s tracking period. Each MOT trial proceeded as described in Fig. 2. Response times (RTs) and accuracy (correct/incorrect) were recorded for DRT trial, and each MOT interrogation (accuracy for the MOT was then summarized as "number correctly classified out of ten").

Results

Prior to analysis, data were excluded using several criteria. We excluded all data from blocks requiring no tracking. The outcome for analysis was tracking performance, and errors in blocks requiring no tracking are likely fundamentally distinct from failures on tracking blocks. We excluded data from four subjects who showed lower than a 70% hit rate on the DRT across all trials (these participants showed hit rates ranging from 29% to 53%). We excluded another seven subjects who failed to respond to the DRT task at all in at least one block. This left 121 (91%) participants in the final analysis. All remaining participants showed DRT hit rates above 80%, with a mean hit rate of 95.85%. The first MOT trial from each block was excluded since we used lagged variables in the following regression analysis. After exclusions, there were 9,772 MOT trials (each corresponding to 15 s of tracking) included in the final analysis. Each trial gave an outcome variable corresponding to the number of dots judged correctly, out of 10. Due to the tendency for errors to occur in pairs we re-expressed the outcome as a proportion out of 5, such that scoring either 8 or 9 out of 10 was coded as 4/5. We suspect this tendency resulted from a combination of (a) the participants knowing the "correct" number of target dots (meaning they were less likely to knowingly make odd-numbers of mistakes), and (b) the nature of the task compounding errors (i.e., mistaking one dot for another would generate two errors – a miss and a false alarm). Our re-coding allows a more parsimonious account of "performance" as a linearly decreasing score.

Descriptive statistics

All descriptive and univariate statistics were performed in Jamovi (The Jamovi Project, 2020), using the "jsq" Bayesian Statistics module (Morey & Rouder, 2018). As expected, RTs to the DRT were slower as tracking load increased (Bayes factor (BF_{10}) > 1,000, Fig. 3). Likewise, hit-rate to the DRT decreased with tracking load (BF_{10} > 1,000, Fig. 3). This replicates our earlier findings that the MOT task can manipulate workload (Howard et al., 2020; Innes et al., 2019), and enables us to confidently use DRT-based statistics as a quantitative measure of "workload."

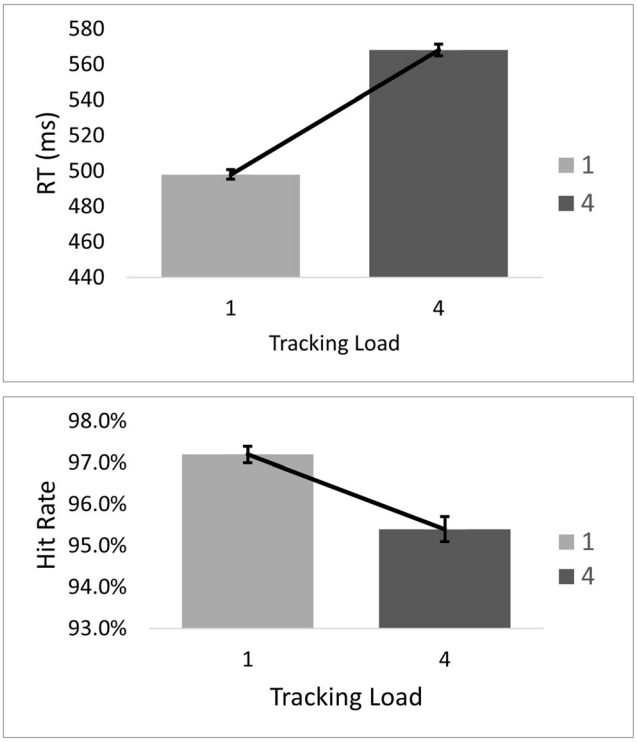

Fig. 3 Response time (**top**) and hit rate (**bottom**) for the detection response task (DRT) task by tracking load in the moving object tracking (MOT) task. In both cases the error bars are the 95% credible interval around the mean

Multi-Level Modeling

To examine the unique predictive relationships between workload (DRT) and prior MOT performance on current MOT performance, we fit a series of multi-level binomial regression models and used model selection to determine the most suitable combination of predictor variables. In each model, the outcome variable was the proportion of MOT dots *correctly classified* as target/non-target expressed as a proportion out of 5. For all models, including the null model, we allowed a random intercept term to capture the base differences in performance across subjects. The models were fit using the "rstanarm" package of R (Goodrich, Gabry, Ali, & Brilleman, 2018; see Online Supplementary Materials for technical details of the model fitting) to allow for Bayesian parameter estimation and inference, and model selection was based on WAIC (Watanabe & Opper, 2010). We fit a total of 32 models allowing for all the possible combinations of five predictors (we provide univariate analyses of each variable in the Online Supplementary Materials):

1. Tracking Load (difficulty; number of dots to track)
2. Current Workload (workload during the trial)

3. Prior Workload (workload over the past 60–90 s, due to random stimulus presentation)
4. Average Workload (subject's mean DRT-RT)
5. Prior Performance (accuracy on the previous trial).

Tracking Load reflects the experimental manipulation (number of dots-to-track). As this manipulation was designed to manipulate workload and difficulty, it was naturally expected to influence performance, and included so that our within-subject workload effects can be interpreted independently from task demands. Current Workload was defined as the average RT on the DRT task within the 15-s period corresponding to the outcome variable. This variable reflects workload *during the time participants were tracking the dots* (see Fig. 4 for a breakdown of subject-level variables). Prior Workload was defined as an exponentially weighted moving average of the RT to the DRT over (up to) the previous 12 DRT trials, weighted toward the most recent trials with a reduction factor of $1/n$. This variable captures workload *prior* to starting the MOT trial, reflecting the possibility for residual workload impacts. The moving average was computed only within block, i.e., trials from blocks with a different Tracking Load were not included in the moving average, and reflects workload across approximately the past 60–90 s (when factoring in time between MOT trials). We used the value of the moving average immediately preceding the start of the outcome MOT trial. Thus, there is no overlap between the two workload measures. Between-subject workload was defined as the mean time to detect the DRT for each subject (including only trials from the one- and four-dot conditions to ensure all subjects were comparable). Prior Performance is the tracking performance (accuracy out of 5) on the MOT trial *preceding* the outcome trial (note that since we removed the first trial from each block *after* lagging, this variable never included performance on a trial from a different Tracking Load).

Predictors 2, 3, and 5 above relate to individual-level measurements. In line with recent suggestions for best practice (Grueber, Nakagawa, Laws, & Jamieson, 2011), we parameterized the coefficients for these predictors such that they could vary across subjects according to a normal distribution centered on each group-level coefficient. This allowed the homogeneity of subject-level workload-performance relationships to be assessed. Once all models had been fit separately, we used WAIC (Watanabe & Opper, 2010) to select the best fitting model, accounting for complexity. The majority of the posterior model evidence (WAIC probability = 0.736) favored the model including all predictors. That is, *tracking load*, *current*, *prior*, and *average workload*, and *performance on the previous trial* all accounted for unique variance in the number of errors on an MOT trial. Each predictor variable showed very strong evidence for inclusion (WAIC Probabilities >0.99), except for the Average Workload (WAIC Probability = 0.738). This suggests that, although all variables were included in the winning model, there is weaker evidence for the inclusion of average workload.

Given the strong posterior evidence for the winning, fully saturated model, we focus exclusively on that model for further analysis. Posterior median group level coefficients are reported in Table 1 with 95% central credible intervals. Tracking Load showed a negative relationship with MOT performance, indicating that as the number of dots-to-track increased from one to four, tracking accuracy decreased (as intended by design). Both within-subject workload variables also show a negative relationship with primary task performance. The coefficient for Current Workload was approximately 1.5 times that of Prior Workload. As these variables are on the same scale, this indicates that workload *during* the performance period is a stronger predictor of performance than workload in the 60–90 s preceding the MOT trial. For a 100-ms increase in DRT-RT (i.e., increased workload) *during the MOT trial*, performance on the subsequent MOT trial is expected to *decrease* by approximately 7.1% (assuming all other variables remain fixed). For Prior Workload, the corresponding decrease would be 4.8%. The subject-level coefficients for Current and Prior Workload were correlated (Pearson's $r = 0.413$, $BF_{10} > 1{,}000$) suggesting participants

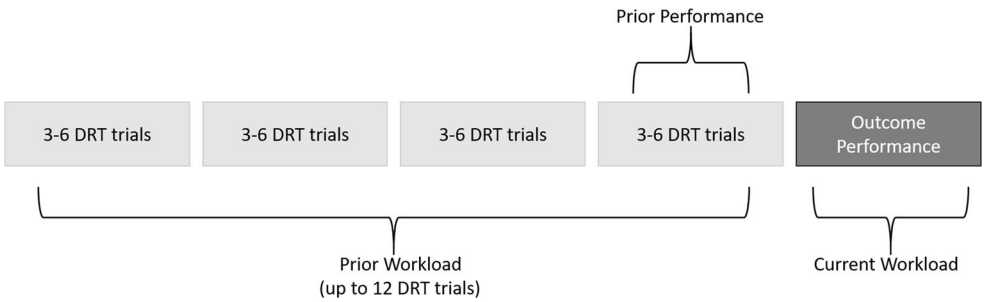

Fig. 4 Graphic representation of the three within-subject variables used in the regression analysis. Each rectangle represents a 15-s moving object tracking (MOT) trial. The outcome variable is performance on the darker "outcome" MOT trial. Current Workload was defined as the detection response task (DRT)-response time (RT) *within the outcome trial*. Prior Workload was defined as an exponentially weighted average of the previous 12 DRT trials (which could cover up to around two to four prior MOT trials depending on the random sequence). Prior performance was defined as the score (out of 5) on the MOT trial prior to the outcome trial

Table 1 Group-level posterior median parameter estimates with 95% central credible intervals

Parameter	Median	2.5%	97.5%
Tracking Load	**-2.041**	**-2.117**	**-1.965**
Current Workload	**-0.738**	**-0.949**	**-0.529**
Prior Workload	**-0.491**	**-0.781**	**-0.207**
Average Workload	**0.820**	**-0.150**	**1.791**
Prior Performance	**0.144**	**0.106**	**0.181**

*Larger absolute values imply a stronger effect (although not all variables are on the same scale). Negative values reflect a negative association (e.g., the negative slope of workload implies performance decreases as workload increases)

who were more impaired by workload during the trial also had stronger "residual workload" effects on performance.

Performance on the previous MOT trial was positively related to performance on the current MOT trial, even when accounting for task difficulty. For every additional dot correctly classified (out of 5) on the previous MOT trial, performance on the following MOT trial is expected to increase by 15%, according to the model coefficient. The coefficient for Average Workload is quite high and, interestingly, positive. This suggests subjects with higher average workloads perform better on the tracking task. However, the credible interval for this parameter is very large, and crosses zero. Therefore, our analysis is inconclusive, but suggests that between-subject effects of workload on performance are likely to be minor at best.

All three subject-level effects were remarkably consistent at the individual-subject level. To reiterate – we allowed each individual participant to have a unique workload->performance (or prior->future performance) slope. We extracted the subject-level parameter estimates for each of the three effects (current and prior workload, and prior performance), and over 95% of subject-level estimates were in the same direction as the corresponding group estimate (this is clear from the abundance of red box-plots in panels 1–2 of Fig. 5, and blue box-plots in panel 3). These parameter estimates were of similar magnitudes. This suggests that the degree and direction of workload-performance decrements is relatively consistent between subjects, and there are few individual differences.

Discussion

Using a sophisticated multi-level modelling approach, we have demonstrated that real-time workload fluctuations at the within-person level are predictive of changes in an individual's own performance. Even when controlling for changes in task load, fluctuations in workload during the time of performance predicted changes in performance on a tracking task. This is consistent with resource trade-off theories (Norman & Bobrow, 1975; Wickens, 2002), as well as recent modelling efforts (Palada et al., 2019). Extending prior findings using subjective workload assessments (Loft et al., 2018; Mracek et al., 2014), we further demonstrate clear residual effects of objectively measured workload on within-person performance on a tracking task. Prior Workload (workload across prior 60–90 s) accounted for unique variance not explained by residual-capacity theories alone. This builds on previous work showing workload and driving performance do not return to baseline immediately after disengaging from a cognitively demanding task (Bowden et al., 2019; Turrill, Coleman, Hopman, Cooper, & Strayer, 2016). The magnitude of current and prior workload effects was correlated at the subject-level, suggesting a generalized susceptibility to workload-based impairment. In a novel contribution, we demonstrated that the directionality of the effects was remarkably consistent (over 95% of subjects showed negative workload-performance relationships). This homogeneity, combined with the correlation between current and prior workload, suggests a generalizable "workload impairment" that is relatively unaffected by individual differences. This knowledge allows more certainty in the conclusion that workload negatively influences performance for most people and that adaptive systems could potentially reliably monitor operator workload in order to adjust task demands.

Performance on the MOT trial immediately preceding the outcome trial was also predictive of performance (also consistent across individuals). This result is particularly intriguing as each trial was entirely independent, so auto-correlation of performance is not guaranteed. Indeed, identifying the detrimental effects of workload and performance *preceding* the outcome trial is the most important contribution of the present study. If variations in workload or performance can be detected early (prior to a task period of interest) and are known to propagate performance impairments forward in time, adjustments could be made to an information display, task scheduling, automation provision, or the division of task-load between operators, to reduce the destructive effect of excess load. This motivates the continued development of "operator-state triggers" for adaptive systems (Feigh et al., 2012; Rouse, 1988). While both past workload and past performance predicted unique variance in tracking performance, it is often true that only one of these variables is known, and systems must therefore adapt as necessary. In some domains, performance is difficult to quantify, but it may be possible to monitor workload, for example by using physiologically derived workload estimates such as heart rate variability or electroencephalography (for reviews, see Charles & Nixon, 2019; Hughes, Hancock, Marlow, Stowers, & Salas, 2019), to drive well-timed work design system adaptations. In practice, however, adaptive systems have proved to be difficult to design,

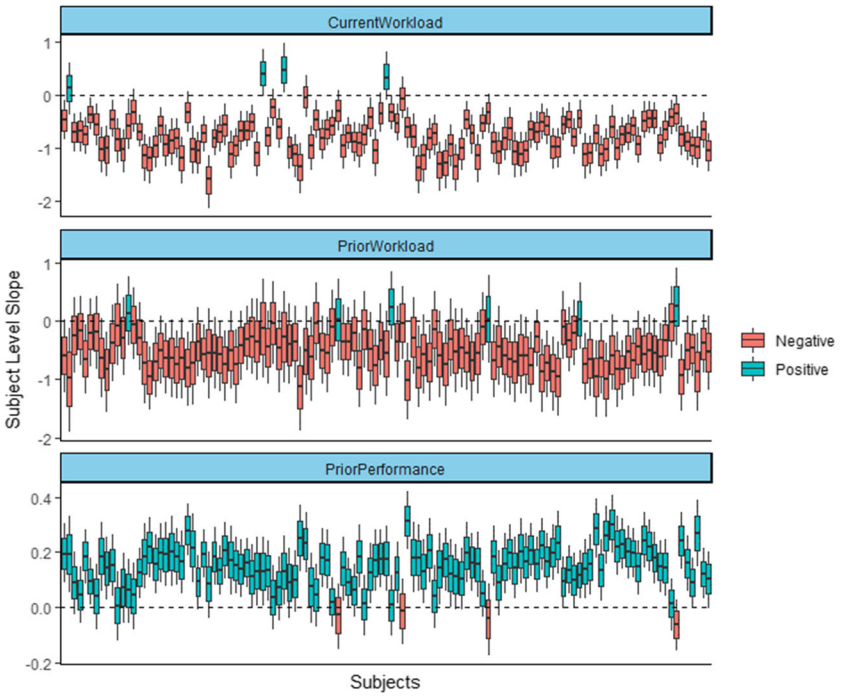

Fig. 5 Posterior estimates of subject-level slopes on current and prior workload, and prior performance. Box-plots reflect the median (bold line), 25% and 75% intervals, with whiskers extending to the 2.5% and 97.5% intervals. Boxes are colored based on the median value (red reflecting a negative association with performance, blue a positive association)

have so-far provided only marginal improvements to performance in the laboratory (e.g., Chen, Visser, Huf, & Loft, 2017; Sauer et al., 2012), and arguably limited operational advantage in real-world contexts. Thus, while our results demonstrate workload can be measured and used to *predict* performance decrements, further work is needed to establish the efficacy of operator-based adaptive systems.

Our results were mixed regarding between-person workload effects (i.e., do between-person variations in capacity influence performance?). Model selection showed that the model including such effects was preferred. However, the positive association of between-subject load and performance in that model was the opposite to that predicted by resource theories. There was also a large amount of variability in the posterior estimates of the between subject parameter, and that variable was associated with weaker evidence than the four other variables. This leads us to a limitation of the study. A key goal was to determine the consistency of within-subject effects. Therefore, we included random slopes on all three of these variables. This means that our five predictor variables were not equally "complex" and dropping only the between-subjects variable would have a negligible influence on model complexity. The complexity added by the within-subject variables also prevented us from exploring potential interactions, which may provide additional insight (e.g., do within-person workload-performance associations manifest more during higher task loads?). As we have now demonstrated the consistency of within-subject workload-performance relationships, future research may be able to explore more complex effects structures by removing the individual variability in regression coefficients.

Another direction that future research should explore is how our results generalize beyond the discrete-trial nature of a laboratory experiment. Although our paradigm was designed to be broadly representative of work contexts such as air traffic control and maritime surveillance, ultimately it is still a discrete, trial-based task. In the real-world, tasks are most often continuous in nature (e.g., air traffic control). In such tasks, performance may not be measured in such binary "correct" or "incorrect" terms, but may instead be expressed as a continuous deviation from some target (e.g., absolute distance from a target altitude in a flight simulator). Real-world tasks often invoke resources from multiple task modalities, which may increase the complexity of the workload->performance relationship due to distinct resource pools (Wickens, 2008). It should also be noted that some theories posit a distinct time-dependent component to the fluctuations in within-person workload (e.g., Grier et al., 2003) and may therefore make different predictions if the time-scale of performance were changed. While our results converge with those of other authors using much longer tasks (e.g., Loft et al. 2018), further research may seek to directly target fatigue and other time-sensitive aspects of workload to assess their unique interactions with task performance.

Ultimately, our study presents novel evidence that increased within-subject workload does indeed impair one's

own performance. Our sophisticated modelling approach allowed us to determine that several types of workload and performance measures account for unique variation in performance on a tracking task. These results have fundamental relevance to applied human factors However, these results are more broadly relevant to the psychological community. We demonstrate clearly that fluctuations in workload *independent from the task demands* accounted for significant variation in performance in a lab task. Our results reinforce the importance of workload as a general construct that should be considered in any study of human cognition and performance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13423-021-01961-6.

Data availability The data and supplementary materials are available on the Open Science Foundation at https://osf.io/6me5z/.

References

Annett, J. (2002). Subjective rating scales: Science or art? *Ergonomics, 45* (14): 966–987.

Baumeister, R. F. (2014). Self-regulation, ego depletion, and inhibition. *Neuropsychologia*, *65*, 313-319.

Boehm-Davis, Deborah A. (Ed); Durso, Francis T. (Ed); Lee, John D. (Ed), (2015). *APA Handbook of Human Systems Integration. APA Handbooks in Psychology*. Washington, DC, US: American Psychological Association, xxix, 625.

Bowden, V. K., Loft, S., Wilson, M. D., Howard, J., & Visser, T. A. (2019). The long road home from distraction: Investigating the time-course of distraction recovery in driving. *Accident Analysis & Prevention*, *124*, 23-32.

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, *74*, 221-232.

Chen, J., Song, X., & Lin, Z. (2016). Revealing the "Invisible Gorilla" in construction: Estimating construction safety through mental workload assessment. *Automation in Construction*, *63*, 173-183.

Chen, S., Visser, T.A.W, Huf, S., & Loft, S. (2017). Optimizing the balance between task automation and human manual control in simulated submarine track management. *Journal of Experimental Psychology: Applied, 23*, 240-262.

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human Factors: 54*(6), 1008-1024.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). rstanarm: Bayesian applied regression modeling via Stan. *R Package Version*, *2*(4), 1758.

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and performance: Volume 2. Cognitive processes and performance* (pp. 41–49). Wiley.

Grier, R. A., Warm, J. S., Dember, W. N., Matthews, G., Galinsky, T. L., Szalma, J. L., & Parasuraman, R. (2003). The vigilance decrement reflects limitations in effortful attention, not mindlessness. *Human Factors*, *45*(3), 349-359.

Grueber, C. E., Nakagawa, S., Laws, R. J., & Jamieson, I. G. (2011). Multimodel inference in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology*, *24*(4), 699-711.

Habib, K., Shalkamy, A., & El-Basyouny, K. (2019). Investigating the effects of mental workload on highway safety. *Transportation Research Record*, *2673*(7), 619-629.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: a meta-analysis. *Psychological Bulletin*, *136*(4), 495.

Hancock, P. A., & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human Factors*, *61*(3), 374-392.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology*, *45*, 73–93.

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*

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: a theory of the relationship between individual differences and information processing. *Psychological Review*, *91*(2), 153.

Hughes, A. M., Hancock, G. M., Marlow, S. L., Stowers, K., & Salas, E. (2019). Cardiac Measures of Cognitive Workload: A Meta-Analysis. *Human Factors, 61*(3), 393–414.

Innes, R. J., Evans, N. J., Howard, Z. L., Eidels, A., & Brown, S. D. (2019). A broader application of the detection response task to cognitive tasks and online environments. *Human Factors*

Innes, R. J., Howard, Z. L., Thorpe, A., Eidels, A., & Brown, S. D. (2020). The effects of increased visual information on cognitive workload in a helicopter simulator. *Human Factors,* 0018720820945409.

Innes, R. J., Kuhne, C. L. (2020) An LBA account of decisions in the multiple object tracking task. *The Quantitative Methods for Psychology, 16(2)*, 175-191.

Kahneman, D. (1973). *Attention and Effort* (Vol. 1063). Prentice-Hall.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *The Behavioral and Brain Sciences*, *36*(6).

Loft, S., Jooste, L., Li, Y. R., Ballard, T., Huf, S., Lipp, O. V., & Visser, T. A. (2018). Using situation awareness and workload to predict performance in submarine track management: A multilevel approach. *Human Factors*, *60*(7), 978-991.

Loft, S., Neal, A., Sanderson, P., & Mooij, M. (2007). Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications. *Human Factors, 49*, 376-399.

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human Factors, 57*, 125–143

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs*. [R package]. Retrieved from https://cran.r-project.org/package=BayesFactor.

Mracek, D. L., Arsenault, M. L., Day, E. A., Hardy III, J. H., & Terry, R. A. (2014). A multilevel approach to relating subjective workload to performance after shifts in task demand. *Human Factors*, *56*(8), 1401-1413.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*(1), 44-64.

Palada, H., Neal, A., Strayer, D., Ballard, T., & Heathcote, A. (2019). Using response time modeling to understand the sources of dual-task interference in a dynamic environment. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(10), 1331.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179-197.

Rouse, W. B. (1988). Adaptive aiding for human/computer control. *Human Factors, 30*, 431–438.

Sauer, J., Kao, C. S., & Wastell, D. (2012). A comparison of adaptive and adaptable automation under different levels of environmental stress. *Ergonomics*, *55*(8), 840-853.

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J. R., Medeiros-Ward, N., & Biondi, F. (2015). Assessing cognitive distraction in the automobile. *Human Factors*, *57*(8), 1300-1324.

The Jamovi Project (2020). *jamovi*. (Version 1.6) [Computer Software]. Retrieved from https://www.jamovi.org.

Thomson, D. R., Besner, D., & Smilek, D. (2015). A resource-control account of sustained attention: Evidence from mind-wandering and vigilance paradigms. *Perspectives on Psychological Science*, *10*(1), 82-96.

Thorpe, A., Nesbitt, K., & Eidels, A. (2020). A Systematic Review of Empirical Measures of Workload Capacity. *ACM Transactions on Applied Perception (TAP)*, *17*(3), 1-26.

Turrill, J., Coleman, J. R., Hopman, R. J., Cooper, J. M., & Strayer, D. L. (2016). The residual costs of multitasking: Causing trouble down the road. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, No. 1, pp. 1967-1970). SAGE Publications.

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(12).

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159-177.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors, 50*(3), 449–455.

Young M. S., Brookhuis K. A., Wickens, C. D., & Hancock P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*, 1–17.

Young, R. A., Hsieh, L., & Seaman, S. (2013). The tactile detection response task: preliminary validation for measuring the attentional effects of cognitive load. In: *Proceedings of the Seventh International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, June 17-20, 2013, Bolton Landing, New York* (pp. 71-77). Public Policy Center: Iowa City, NY. University of Iowa.