Cognitive Workload Measurement and Modeling Under Divided Attention

Spencer C. Castro[1]
David L. Strayer[1]
Dora Matzke[2]
Andrew Heathcote[3]

Author Note:

1, Department of Psychology, University of Utah

2, Department of Psychology, University of Amsterdam

3, Division of Psychology, University of Tasmania

This research was supported in part by the National Science Foundation Graduate Research Fellowship Program, the AAA Foundation for Traffic Safety, the Veni grant (451-15-010) from the Netherlands Organization for Scientific Research (NWO), ARC DP160101891, and CERA247.

Word Count: 9088

Correspondence concerning this article should be addressed to Spencer Castro, Department of Psychology, University of Utah, Salt Lake City, Utah, 84112

Email contact: spencer.castro@psych.utah.edu

Abstract

Motorists often engage in secondary tasks unrelated to driving that increase cognitive workload, resulting in fatal crashes and injuries. An International Standards Organization (ISO) method for measuring a driver's cognitive workload, the Detection Response Task (DRT), correlates well with driving outcomes, but investigation of its putative theoretical basis in terms of finite attention capacity remains limited. We address this knowledge gap using evidence-accumulation modeling of simple and choice versions of the DRT in a driving scenario. Our experiments demonstrate how dual-task load affects the parameters of evidence-accumulation models. We found that the cognitive workload induced by a secondary task (counting backward by threes) reduced the rate of evidence accumulation, consistent with rates being sensitive to limited-capacity attention. We also found a compensatory increase in the amount of evidence required for a response and a small speeding in the time for non-decision processes. The ISO version of the DRT was found to be most sensitive to cognitive workload. A Wald-distributed evidence-accumulation model augmented with a parameter measuring response omissions provided a parsimonious measure of the underlying causes of cognitive workload in this task. This work demonstrates that evidence-accumulation modeling can accurately represent data produced by cognitive workload measurements, reproduce the data through simulation, and provide supporting evidence for the cognitive processes underlying cognitive workload. Our results provide converging evidence that the DRT method is sensitive to dynamic fluctuations in limited-capacity attention.

Public Significance Statement: People around the world endanger the lives of themselves and others every day by dividing their attention across multiple tasks, such as driving and talking on a cell phone. These dangers result from splitting and overtaxing our limited voluntary attentional efforts. Current tools for measuring attentional effort, also known as cognitive workload, lack insight into cognitive factors that can cause fatal errors. With the advent of new distracting technology in cars, if we do not effectively measure cognitive workload fatal human errors may grow. To quantify cognitive workload under a simulated driving-like task, the current study details our application of mathematical modeling to an International Standard for measuring ongoing cognitive workload in the vehicle. This research provides a framework for accurately quantifying cognitive workload and the factors that contribute to it, which will allow future researchers and policy makers to determine the danger inherent in many tasks within the vehicle.

Keywords: Detection Response Task, Driving Simulation, Wald Distribution, Independent Race Model, Cognitive Workload, Evidence Accumulation Modeling, Attention, Human Performance, Multitasking, Dual Task Cost

The capacity limits of human cognition play a central role in performing everyday activities. These limits of capacity affect psychological constructs from self-regulation (e.g., Muraven & Baumeister, 2000) to the prevalence of stereotyping (e.g., Biernat, Kobrynowicz, & Weber, 2003), but they become most apparent under divided attention: when people attempt to perform more than one cognitive activity at the same time (e.g., driving an automobile and using a smartphone). Performing these tasks requires cognitive effort, or a *mental workload* that must be maintained to achieve the concurrent goals. Hart and Staveland (1988) define *mental workload* as the "…relationship between the amount of mental processing capability or resources and the amount required by the task". Based on this view of workload, Strayer, Watson, and Drews (2011) emphasize *cognitive* sources of distraction to distinguish them from *visual* and *manual* components, which all contribute to overall workload. They argue that divided attention, such as when driving and talking on a cell phone, decreases performance in both tasks largely due to the cognitive component of workload.

Although robust, the precise cause of cognitive workload's effect on performance under divided attention is less well understood. On the one-hand, declines in performance may stem from reductions in the *efficiency* of information processing, perhaps due to a competition for a limited pool of resource (i.e., cognitive capacity). On the other hand, declines in performance may reflect a bias to more conservative responding under higher cognitive workload (i.e., response caution). Though it is often difficult to distinguish between these alternative interpretations, this distinction has important implications for theories of attention in complex multitasking situations. In the former, the *rate* of information processing is slowed by multitasking. In the latter, the *threshold amount* of information required for decisions is

increased by multitasking. Given the ubiquity of multitasking in modern society, this distinction also has important real-world consequences.

For example, the National Highway Transportation Safety Administration (NHTSA) found that at any given time over 10% of drivers are using a cellular device (NHTSA, 2016). Although NHTSA (2012, 2016) guidelines currently cover only visual and manual sources of distraction, Klauer et al. (2014) demonstrated that deficits in attention—largely caused by the cognitive workload required to drive and perform secondary tasks such as using a mobile device—are a leading factor in the majority of crashes and near-crashes. Castro, (2017) demonstrated that mobile devices of different sizes, whether they are handheld or mounted, differentially affect attention to changes beyond the device. Depending upon the specific mechanisms underlying cognitive workload's impact on crash risk, studies measuring cognitive workload may recommend different solutions to ameliorate these risks. If cell phone use impacts the rate of information a driver is capable of processing, then strategies and policies that optimize a driver's allocation of limited resources to the road are warranted. However, if drivers change their behavior by requiring more information from their environment before making decisions, then perhaps updated driver training may be recommended to decrease decision time. Of the two outcomes, previous research would seem to support theories of limited resources and deficits resulting from the rate of information processing; however, this assumption has yet to be rigorously evaluated.

Theoretical models of human psychological-performance limitations stem largely from research on the role of attention in goal-directed behavior. Kahneman (1973) describes voluntary, goal-directed attention as a finite capacity that limits information processing speed. In resource theories (e.g., Navon & Gopher, 1979), capacity is shared among tasks operating in parallel, with the processing rate for each proportional to its attention allocation. In single-

channel bottleneck theories (e.g., Welford, 1952) attention is interleaved, switching in an all-or-none manner among tasks, with each task's average processing rate proportional to the attention it receives. This theory has been applied to driving utilizing Welford's (1952) Psychological Refractory Period (PRP) paradigm. Levy, Pashler, and Boer (2006) demonstrated that the PRP is evident in the driving context by demonstrating slowed brake reaction times (RTs) when occurring shortly after an auditory discrimination stimulus. The researchers varied the stimulus-onset asynchrony (SOA) between the auditory stimulus and requiring the participant to brake, showing that brake RT increased with shorter SOAs. This paradigm demonstrates that dual-task performance leads to serial processing of discrete stimuli and responses. However, it has a somewhat limited application to driving, which can be categorized as a slightly automated, cognitively demanding continuous task. The potentially distracting secondary tasks provide intermittent dual tasking, but they occur in parallel with the primary continuous task of driving. PRP designs require two discrete responses and do not seem to be the best candidate for measuring ongoing driving performance decrements induced by cognitive workload. Their secondary tone task should be sufficient to predict any driving performance issues and would be sufficient with the approach outlined in this paper. In both theories, attention-degrading secondary tasks produce a load that detracts from primary-task performance. Strayer and Fisher (2016) argue that load induced by cognitive sources of distraction in driving account for failures to notice objects in the fovea (Strayer & Drews, 2007), increased brake reaction time (Caird, Willness, Steel, & Scialfa, 2008), failures to stop at intersections (Strayer, Watson, & Drews, 2011), and decreased visual scanning (Taylor et al., 2013).

The Detection Response Task (DRT) was developed by the International Standards Organization (ISO) to measure the potentially lethal and difficult to quantify cognitive workload effects of secondary tasks (ISO DIS 17488, 2015). It provides a simple measurement of cognitive

workload that directly correlates with good driving performance (Strayer et al., 2015), retrospective subjective workload measures (Hart & Staveland, 1988), and electrophysiology (Strayer et al., 2015). The DRT requires a button press in response to an easily-detected stimulus that occurs randomly every 3-5 seconds. Increased secondary-task load causes slowing in DRT response time and/or an increased response omission rate (Castro, Cooper, & Strayer, 2016; Cooper, Castro, & Strayer, 2016; Strayer, Biondi, & Cooper, 2017). Recently, the NHTSA has taken note of the DRT's efficacy and practicality and plans to incorporate it into their Driver Distraction Guidelines (Ranney, Baldwin, Smith, Mazzae, & Pierce, 2014).

ISO DIS 17488 (2015) claims that with appropriate apparatus (i.e., stimuli and responses that do not overlap with other tasks) the DRT has only minimal effects on driving. The typical DRT manipulation employed by researchers consists of baseline driving (driving + DRT) and then driving with a secondary task (driving + DRT + secondary task). This experimental design enables comparisons that quantify the cognitive workload of the secondary task but do not directly address the DRT's impact on driving performance. Previous studies have found mixed results for the effect of the DRT alone on driving (e.g., Ranney, Baldwin, Smith, Mazzae, & Pierce, 2014; Strayer et al., 2015). Castro, Cooper, and Strayer (2016) developed a simulated steering task that allows more sensitive measurement and concluded that there is a small effect of the DRT. Although this effect is usually inconsequential for real driving performance, it is theoretically important, as it is consistent with the idea of capacity sharing between the DRT and other tasks.

Even though it is increasingly adopted as a standard for making critical judgments, such as deciding how to instrument cars and rank their safety (Strayer et al., 2015), validation of the DRT has been mostly empirical, with only a few investigations of its theoretical underpinnings (Ratcliff & Strayer, 2014; Tillman, Strayer, Eidels, & Heathcote, 2017). Given the importance of

assessing cognitive workload and the allocation of attention in a wide variety of dynamic environments, understanding what the DRT is measuring is important in both basic and applied contexts. We expanded this line of research using the DRT and evidence-accumulation modeling. Evidence-accumulation modeling is a theoretical framework that has been successfully applied to understand speeded responding in a wide range of choice tasks that require the selection of a set of two or more response options (Brown & Heathcote, 2008; Leite & Ratcliff, 2010), and less widely to simple tasks like the DRT requiring only one response (e.g., Ratcliff, 2015). For both simple and choice tasks, this framework assumes an initial encoding stage that extracts evidence from a stimulus. Next, an accumulation stage accrues evidence until it reaches a threshold amount, at which point a final response-production stage is initiated. Response time (RT) equals the time to reach threshold (decision time) plus non-decision time, which is the sum of encoding and response production times.

Evidence-accumulation modeling allows for a more fine-grained representation of the mechanisms underlying the impacts of secondary tasks on driving, and thus can improve the validity of DRT studies. In particular, these models incorporate parameters representing the rate of information processing (i.e., drift rates), lower level perceptual and motor influences (i.e., non-decision times) and higher order strategic caution and bias processes (i.e., response thresholds). A variety of evidence-accumulation models have been proposed that although sharing a set of core assumptions and parameters, differ in some details. We employed two different models, the LBA (Brown & Heathcote, 2008) and Wald (Leite & Ratcliff, 2010), in order to check if these differences in detail influenced the inferences we made about underlying mechanisms based on their core parameters. We found that both models led to the same conclusions and we focus on the Wald model here. It provided a slightly better model of our data

and it has been more often used with the DRT, so it better supports comparisons to previous results. LBA results are reported in Supplementary Materials.

Figure 1 illustrates an evidence-accumulation model of the DRT, where the evidence total (dashed line) is stochastic (i.e., varying from moment-to-moment) and increasing at a mean rate *v* towards a threshold *b*. This is called a Wald" model because, assuming infinitesimal Gaussian moment-to-moment variability, the distribution of decision times follows a positively skewed Wald distribution, which in combination with a shift (non-decision time) parameter provides a good description of simple RT (Heathcote, 2004). It can be extended to model choice tasks by having an accumulator independently gathering evidence for each option, with the first to reach its threshold causing the corresponding response to be selected (Leite & Ratcliff, 2010).

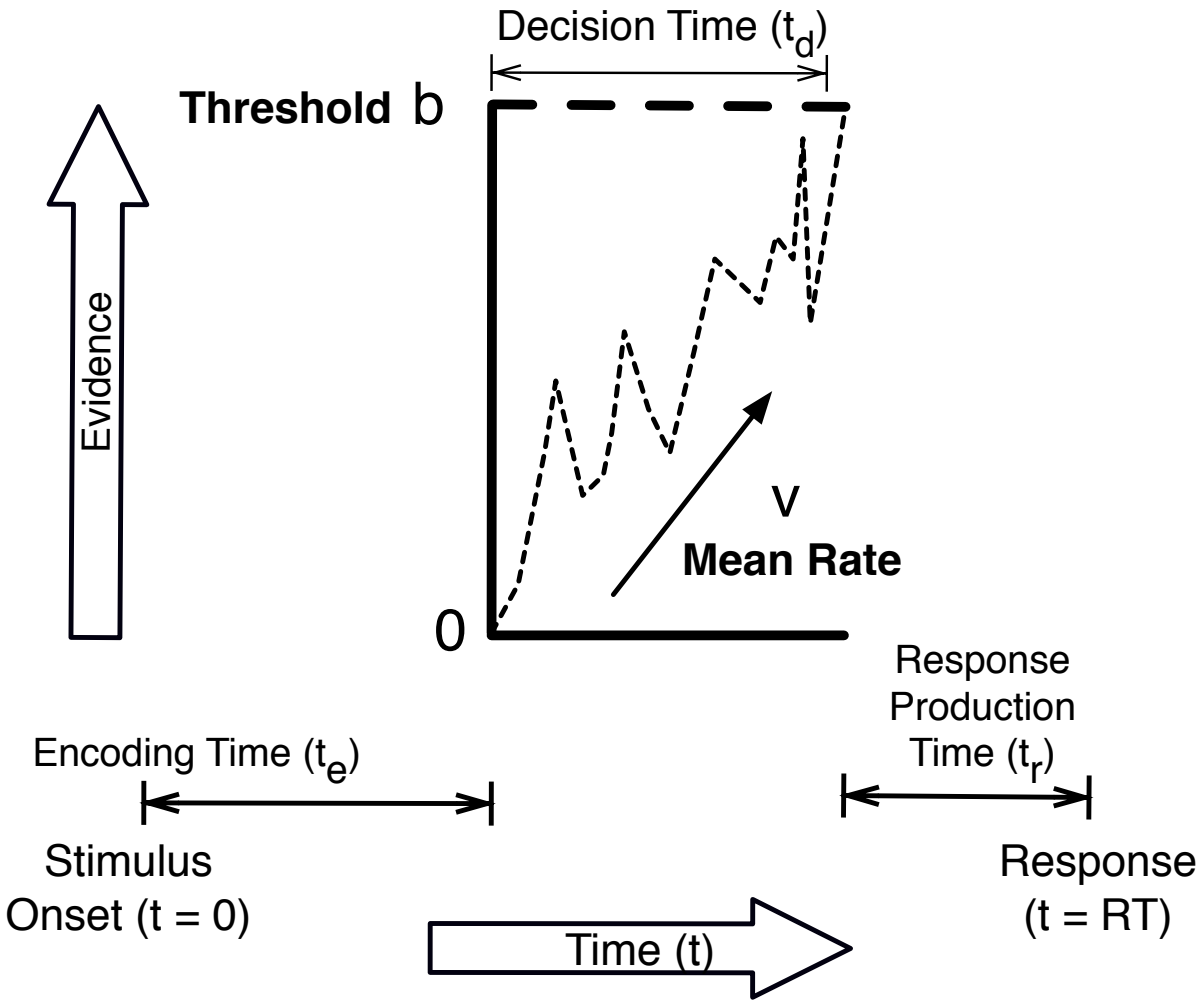

Figure 1. The Wald evidence-accumulation model for the DRT. Note the dashed evidence accumulation path is a caricature and would in reality vary more rapidly.

Given that cognitive workload is thought to affect the rate of information processing, it naturally maps to evidence-accumulation rate parameters. However, Heathcote, Loft, and Remington (2015) demonstrated in the domain of prospective memory—which was thought to slow performance of a primary or ongoing task because of capacity sharing with a monitoring process required to achieve a prospective-memory goal—that slowing of task performance stemmed primarily from individuals delaying ongoing responses to make it less likely that they pre-empted the response required by the prospective-memory goal. This conclusion called into

question prevailing limited-capacity theories of prospective memory (see also Strickland, Loft, Remington & Heathcote, 2017, in press).

In the domain of cognitive workload, it is possible that effects are mediated by threshold changes and/or changes in a number of other factors. Increased cognitive workload may slow early perceptual encoding, and hence non-decision time, or even cause failures to encode evidence from the stimulus, and hence response omissions. Previous studies of perceptual choice tasks claim that visual-attention-load effects are reflected in accumulation rates (Eidels, Donkin, Brown, & Heathcote, 2010). Schmiedek et al. (2007) correlated individual differences on a variety of tasks (e.g., working memory, reasoning, and psychometric speed) and evidence-accumulation rates in verbal, numeric and spatial choice paradigms as reflecting attention capacity as well. Cognitive workload may also affect the quality of evidence, which is inversely proportional to the level of noise in evidence; in choice tasks it is directly proportional to the difference between evidence for the different options. Lower quality evidence can result in choice errors unless evidence is collected for a longer time, and so may indirectly cause slowing if participants raise their threshold to maintain accuracy (i.e., a "speed-accuracy tradeoff).

In tasks like the DRT where there is only one response, higher noise in evidence associated with reduced attention may also cause false detection responses (e.g., due to moment-to-moment fluctuations that occur even when no stimulus is present), and so again participants may have to set larger thresholds with the increased cognitive workload. Alternatively, higher thresholds may reflect a general tendency to be more cautious when making responses in more demanding situations (Strickland et al., 2017; Tillman et al., 2017). Consequently, it is an open question as to which aspects of information processing are altered under divided attention and exactly what aspects of information processing are being measured by the DRT methodology.

Experiment

Cognitive workload effects are prototypically measured in dual-task paradigms, by contrasting primary-task performance between conditions with and without a secondary task that is attention demanding, but which does not overlap with the primary task regarding perceptual and motor components. We performed such a dual-task experiment with eight conditions (see Table 1), with a driving-like primary task (Castro et al.'s 2016 steering task) common to all. In four of the eight conditions participants also performed a secondary task of counting backward by threes. The four conditions with a secondary task and four conditions without a secondary task differed in the same way with respect to requirements related to the DRT. In a baseline condition there was no DRT, and so no cognitive workload measurement was taken. In a second condition, cognitive workload was measured with a typical ISO DRT to a bright light. In a third condition, the DRT used a dimmer light, and in a fourth condition cognitive workload was measured with a choice version of the DRT, where participants had to press one of two buttons to indicate whether the light was dim or bright. We hypothesized that the secondary-task load would slow DRT responding and increase errors (i.e., missed responses and also incorrect responses in the choice DRT). From past work on visual workload and individual differences, we hypothesized that the secondary task would reduce evidence accumulation rates in the DRT, and also increase thresholds and non-decision times.

Method

Participants

After Institutional Review Board approval, twenty participants (17-28 years old, M=20.2) were recruited via psychology courses at the University of Utah (10 males, 10 females) and were compensated for class credit upon completion of two two-hour sessions on different days. All reported normal visual acuity and normal color vision. Each participant completed a

large number of trials per condition as this, and not the number of participants, was critical for parameter estimation (see supplementary materials for parameter recovery, showing that we had adequate sample size and data quality).

Materials

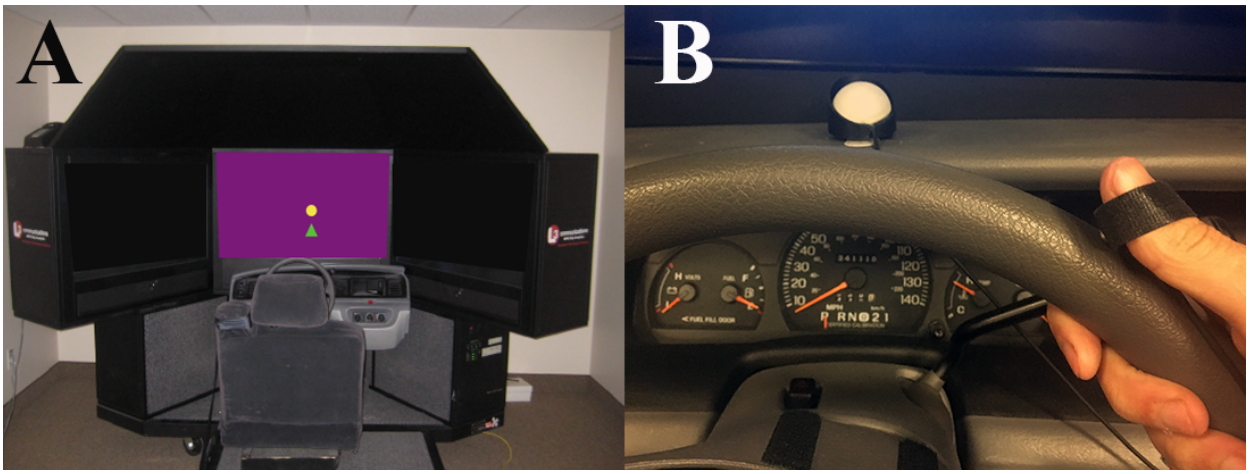

Figure 2. Photograph A shows the simulator used to display the pursuit-tracking task utilizing the steering wheel and center screen. Participants control the triangle in an attempt to keep its lateral position equivalent to the circle's lateral position. Photograph B shows the dash-mounted DRT for displaying the dim and bright simple DRT, and choice DRT, stimuli.

A 101.6 cm Samsung LCD (1920 x 1080 pixels) was used to display the pursuit-tracking task (see Figure 2). Participants utilized a steering wheel from a driving simulator to track a ball that moved continuously on the screen with a triangle cursor (see Figure 2A). The ball had a diameter of 20 pixels (~0.96 cm), which was the same length as the sides of the equilateral triangle cursor. The steering wheel updated the location of the cursor through a Sparkfun™ Electronics rotary encoder set to sample the position at 30 Hz. The DRT device presented a dash-mounted light at two intensities of red (see Figure 2B). Stimuli were presented randomly every 3-5 seconds and responses were made by pressing one of two micro-switches attached to participants' left and right thumbs.

Design

The pursuit-tracking task was created to simulate steering on a moderately curvy road. Participants were instructed to maintain the cursor as close as possible to a ball that moved horizontally across the screen at a slow constant rate of 100 pixels per second (see Figure 2A). As the ball approached the edge of the screen, it became more probable that the ball would reverse direction and maintain its constant movement in the other direction. The probability of the ball's location followed a normal distribution centered on the middle of the screen, so that, for example, the ball moved smoothly through the center third of the screen (corresponding to one standard deviation either side of the middle) approximately 68% of the time, and the center two thirds approximately 95% of the time (corresponding to two standard deviations either side of the middle).

There were four pursuit tracking conditions: single-task tracking (i.e., DRT absent), and tracking while concurrently making a detection response to the onset of a low-intensity (i.e., dim) light, a high-intensity (i.e., bright) light, or a choice response to a dim vs. bright light (see Table 1). These conditions were crossed fully with a cognitive workload (i.e., load) manipulation of either counting backward by threes (i.e., load present) or not counting (i.e., load absent; see Table 1). The 4 x 2 factorial design was blocked into 64 one-minute runs and counterbalanced using a balanced Latin-Square design. Participants were given 30 seconds of rest between each block. Apart from single-task tracking (i.e., DRT absent), there were an average of 240 DRT trials for each cell of the design.

Table 1

Experimental Design with Factors DRT Stimulus Type (4) x Cognitive Workload (2).

Cognitive Workload	DRT Type			
	DRT Absent	Simple		Choice
		Bright Light	Dim Light	
Absent	DRT Absent Load Absent	DRT Bright Load Absent	DRT Dim Load Absent	DRT Choice Load Absent
Present	DRT Absent Load Present	DRT Bright Load Present	DRT Dim Load Present	DRT Choice Load Present

Choice Difficulty Calibration

Before the experiment, the lights were calibrated so that each participant was approximately 75% accurate in their choice classification based on calibration algorithms proposed in Macmillan and Creelman (2005). The ISO DRT has a brightness range for its light emitting diodes (LEDs) from 0 (off) to 255 (brightest) (ISO DIS 17488, 2015). We initially set the values of the bright and dim lights to 200 and 100 respectively. Participants made sets of 8 choice responses; then the dim light was changed by the proportion correct multiplied by a weight that decreased for each set of 8 responses from 150% toward 0% in progressively smaller amounts (see Figure 3). When participants scored below 75% the light difference was increased by the weighted amount. Participants proceeded to the main experiment when 75% accuracy was achieved for three consecutive blocks, with light intensities after that remaining fixed.

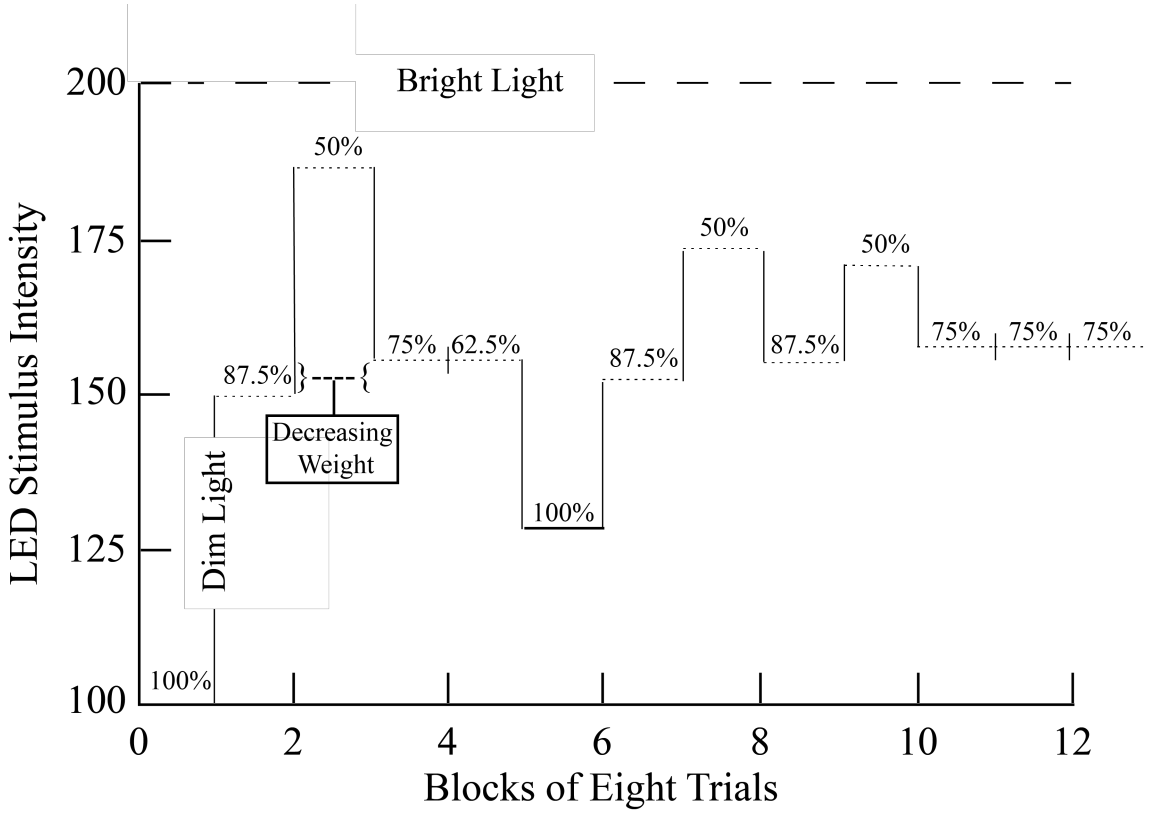

Figure 3. Following Figure 11.6 of Macmillan and Creelman (2005), a calibration procedure of 13 steps for a hypothetical participant. The weight decreases according to the inverse power law $f(x) = Cx^{-1/2}$, where *C* is a constant truncating the weight at 150%, and *x* corresponds to the step number. The percentages refer to the accuracy averaged over 8 trials.

Measures

RT to the dashboard light was recorded to the nearest millisecond. RTs shorter than 150 milliseconds and trials with two or more responses were excluded from the analyses (0.78%). Also, blocks with fewer than 8 presented DRT trials over the course of a 1-minute block were removed (1.2%), as were blocks with lower than 50% accuracy, or lower than 50% responses (0.76%, 0.09%, respectively). The Root Mean Squared Error (RMSE) of the pursuit-tracking task was computed from differences between the position of the cursor and the target sampled at 30 Hz. The RMSE was calculated with the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\overline{y_i} - y_i)^2}{n}}$$

where the sum is of 1 to *n* observations (i.e., ~1780) taken over the course of an ~1-minute block of the lateral position of the cursor ($\overline{y}$) minus the position of the target (y) in each 30th of a second interval. The pursuit-tracking task failed to record for three participants, resulting in a loss of data. Any RMSE tracking error recorded 3 standard deviations above the individual participant's mean was also removed (1.20%).

Results

All analyses used R (R Development Core Team, 2016)[1]. We first report conventional analyses of tracking error, the proportion of omissions, accuracy and mean RT using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015). Table 2 describes which measures were available in each condition. Participants were included as a random effect, and we used Type II Wald chi-square tests. We report 95% confidence intervals in square brackets. Table 3 contains a summary of omissions, accuracy, and RT-mean comparisons.

Table 2

Presence (+) of dependent variables for the different DRT Types.

Dependent Variable	DRT Type			
	DRT Absent	Simple		Choice
		Bright Light	Dim Light	
Pursuit Tracking	+	+	+	+
Omissions		+	+	+
Response Time		+	+	+
Choice Accuracy				+

[1] We have provided our dataset, analyses, and our custom software as a public project on the Open Science Framework (https://osf.io/e8kag/). We have provided a template for future modeling of cognitive workload measurements with the models utilized within this repository.

Pursuit Tracking Measures

Cognitive Workload. Collapsing across DRT types, RMSE steering error was greater for the load-present condition (M = 2.23, [2.21, 2.24]) than the load-absent condition (M = 2.16, [2.14, 2.17]), $\chi^2(1) = 157.92$, $p < .001$.

DRT Type. Collapsing across the cognitive workload manipulation, we performed pairwise comparisons of the four DRT Types to their closest performer. The bright (i.e., ISO standard) DRT increased steering error (M = 2.15, [2.14, 2.17]) over the single-task condition (i.e., DRT absent; M = 1.97, [1.96, 1.99]), $t(16) = 3.83$, p = .001, [.08, .28], but had a significantly smaller steering error than the dim DRT condition (M = 2.22, [2.20, 2.23]), $t(16) = 4.58$, $p < .001$, [.03, .09]). The choice DRT (M = 2.38, [2.36, 2.39]) significantly increased steering error over the dim DRT condition as well, $t(16) = 3.65$, p = .002, [.06, .24].

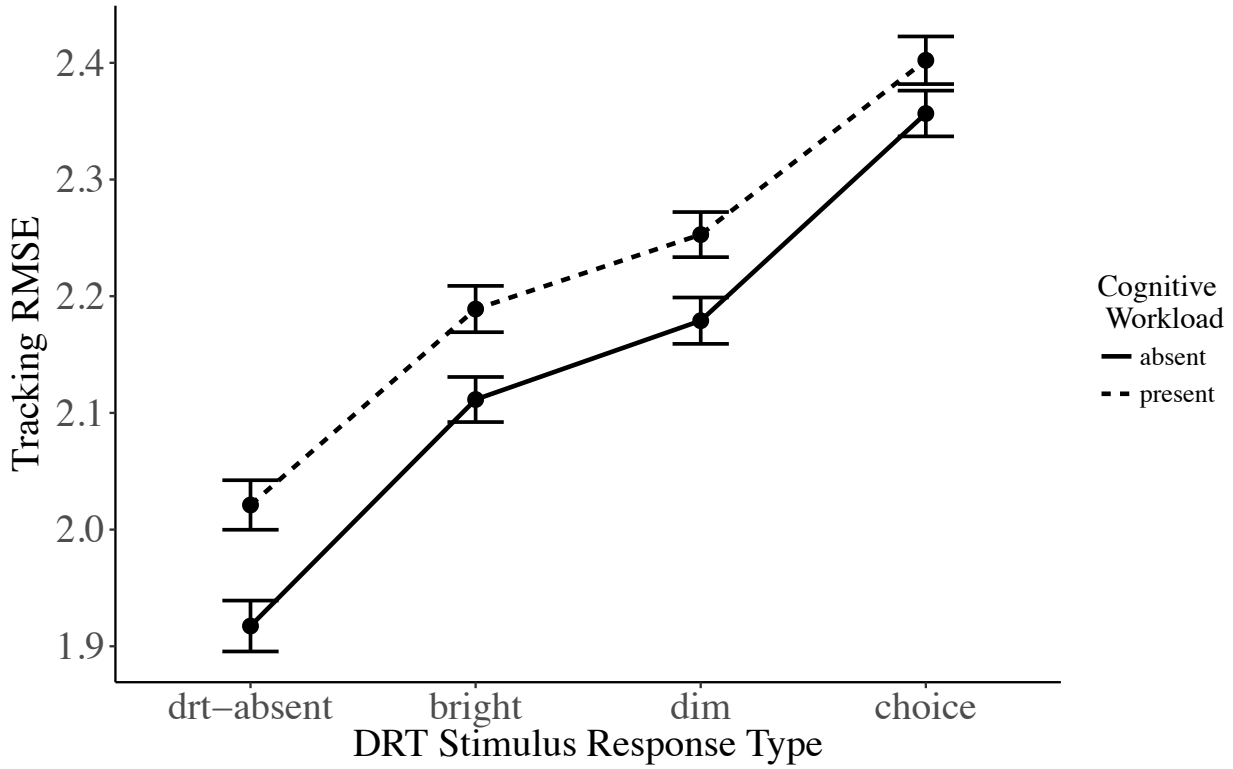

Figure 4. Root Mean Squared Error (RMSE) for the tracking task. Error bars are 95% confidence intervals around the mean utilizing the Cousineau-Morey method (Cousineau, 2005; Morey, 2008; Baguley, 2012).

Additionally, the average size of the load effect across the simple DRTs (i.e., bright and dim DRTs) did not differ significantly from the load effect in the DRT-absent condition $t(16) = 0.75$, $p = .46$, [-.09, .22], but the load effect with the addition of the choice task was significantly smaller $t(16) = 2.80$, $p = .005$, [.07, .28], driving an interaction between load and the addition of different DRTs $\chi^2(3) = 14.75$, $p = .002$ [.02, .10] (see Figure 4).

DRT Measures

DRT measures came from the bright and dim simple DRT, and choice DRT types. For our analyses, these DRT types were sometimes grouped as simple (bright, dim) vs. choice. These measures include the percent of omissions, (i.e., failures to respond within 3 seconds after stimulus presentation), RT, and the percent of correct light discriminations in the case of the choice task (see Table 2).

Omissions. We combined data from the bright and dim DRTs to make a 2 (DRT type) × 2 (cognitive workload) design. For the simple DRTs, omission rates were lower when in the load-absent condition ($M = 4.30\%$, [3.70, 4.89]) than the load-present condition ($M = 5.96\%$, [5.24, 6.67]), $\chi^2(1) = 18.66$, $p < .001$. Neither the effect of stimulus type, nor its interaction with load, was significant. For the same design – but replacing the simple conditions with the two-alternative choice task – load also significantly affected omission rates, $\chi^2(1) = 36.64$, $p < .001$. Participants failed to respond more often in the load-present condition ($M = 3.60\%$, [2.73, 4.47]) than when in the load-absent condition ($M = 1.31\%$, [.80, 1.83]), but neither the effect of stimulus type, nor its interaction with load, was significant. The increase in omissions due to load was significantly greater for the choice (2.00%, [1.57, 3.01]) than the simple DRT (1.10%, [.96, 2.36]), $\chi^2(1) = 10.62$, $p = .001$ (see Figure 5).

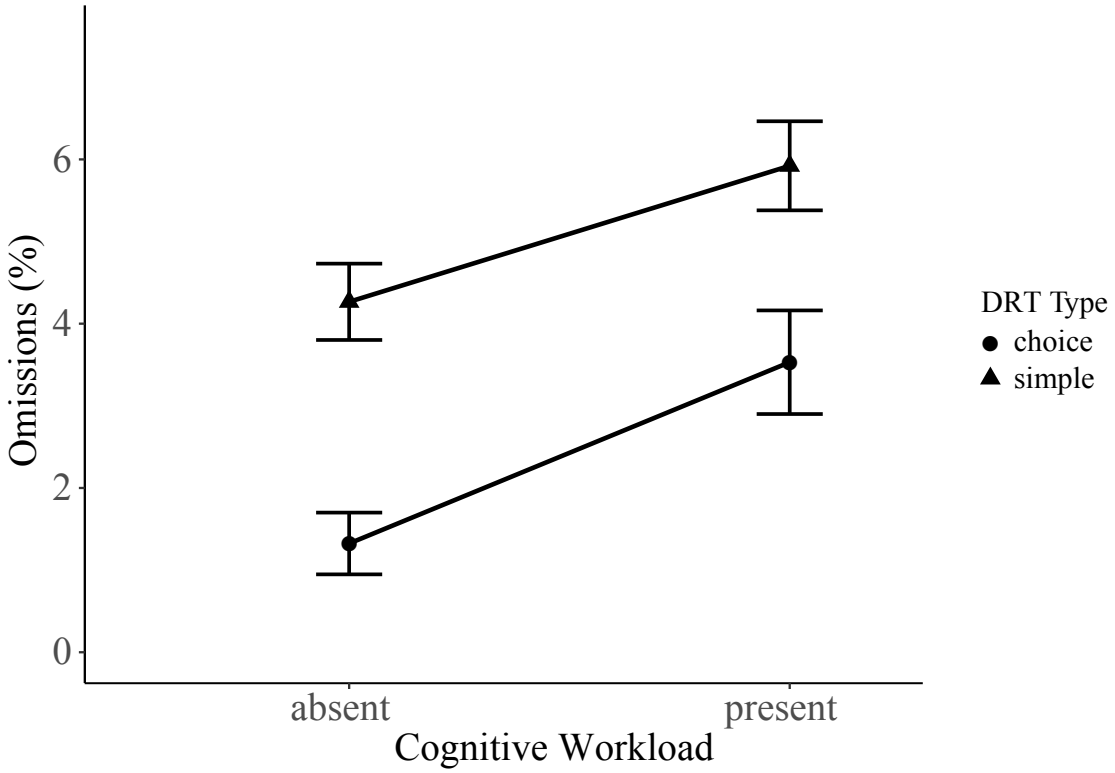

Figure 5. Percent omissions (failures to respond) to DRT stimuli. The size of the cognitive workload effect differed between the simple DRTs (i.e., Bright and Dim) and the choice DRT. Error bars are 95% confidence intervals around the mean utilizing the Cousineau-Morey method (Cousineau, 2005; Morey, 2008; Baguley, 2012).

Response Time. We again combined data from the two simple DRTs to make a 2×2 design and transformed the RT data to the log scale for analysis, but report means on the seconds scale (see Figure 6). Participants responded 0.146 s slower in the load-present condition (M = 0.622 s, [.614, .631]) than the load-absent condition (M = 0.479 s, [.472, .483]), $\chi^2(1) = 1711.95$, $p < .01$. The main effect of stimulus was also significant, $\chi^2(1) = 37.43$, $p < .001$, but participants were only 0.018 s slower for the dim stimulus (M = 0.555 s, [.546, .561]) compared to the bright stimulus (M = 0.539 s, [.529, .543]). The two effects did not interact significantly.

In the choice DRT, participants responded 0.098 s slower in the load-present condition (M = .966 s, [.951, .980]) than in the load-absent condition (M = .868 s, [.856, .881]), $\chi^2(1)$ =

230.49, $p < .001$. The main effect of stimulus was again significant, $\chi^2(1) = 5.64$, $p = .018$, but small—0.019 s slower for the dim stimulus ($M = .923$ s, [.909, .937]) compared to the bright stimulus ($M = .904$ s, [.891, .917])—and again the two effects did not interact significantly. We also found that participants were 0.049 s slower overall on error trials ($M = .951$ s, [.929, .973]) compared to correct trials ($M = .902$ s, [.891, .912]), $\chi^2(1) = 34.04$, $p < .001$.

Overall, the simple DRT was much quicker than the choice DRT ($M = .545$ s, [.540, .550] vs .913 s, [.904 0.923]), $\chi^2(1) = 212.00$, $p < .001$, and the increase in mean response time due to load was significantly greater for the simple ($M = .151$ s, [.140, .162]) than the choice ($M = .102$ s, [.088, .116]) DRT, $\chi^2(1) = 212.13$, $p < .001$.

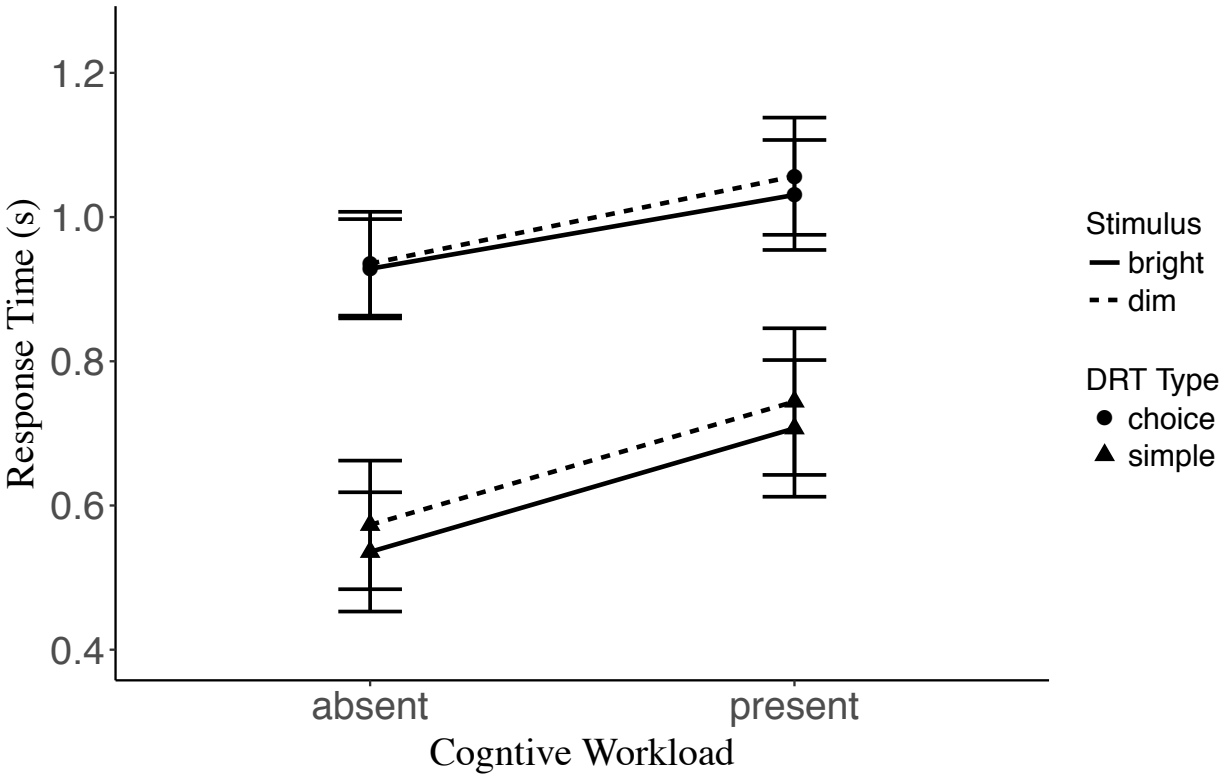

Figure 6. Response time for the average of simple DRT conditions and the choice DRT. Error bars are 95% confidence intervals around the mean utilizing the Cousineau-Morey method (Cousineau, 2005; Morey, 2008; Baguley, 2012).

Choice Accuracy. Participants were more accurate in the load-absent condition ($M =$ 77.40%, [75.48, 79.27]) than in the load-present condition ($M = 74.60\%$, [72.15, 76.70]), $\chi^2(1) =$

9.28, p = .002), but neither the effect of stimulus type, nor its interaction with the load effect, was significant (see Table 3 for a summary).

Table 3

Means and Standard Deviations for stimulus type (2) by Cognitive Workload (2) for the simple and choice tasks.

Task	Dependent Variable	Stimulus or Load	Mean	SD	*p*
Simple	Omissions (%)	Absent-Present	4.6%, 6.2%	9.4%, 8.0%	<.01
		Bright-Dim	5.4%, 5.3%	8.8%, 8.5%	.40
	Response Time (s)	Absent-Present	.48 s, .62 s	.19 s, .21 s	<.01
		Bright-Dim	.54 s, .56 s	.18 s, .20 s	<.01
Choice	Omissions (%)	Absent-Present	1.3%, 3.6%	1.1%, 3.4%	<.01
		Bright-Dim	2.1%, 2.7%	1.8%, 2.4%	.04
	Response Time (s)	Absent-Present	.87 s, .97 s	.15 s, .17 s	<.01
		Bright-Dim	.90 s, .92 s	.16 s, .15 s	.02
	Accuracy (%)	Absent-Present	77.4%, 74.6%	6.2%, 6.5%	<.01
		Bright-Dim	76.9%, 75.2%	11.8%, 8.8%	.08

Discussion

In summary, the RMSE tracking error and RT measures exhibited an increase with cognitive workload and DRT difficulty, but the cognitive workload difference was smaller for the choice DRT. With the addition of the choice DRT, both responses to the DRT and RMSE tracking error were less affected by the cognitive workload manipulation. Accuracy in the choice DRT also decreased with load.

We now use evidence-accumulation modeling to understand the underlying causes of the behavior we observed. We first introduce the model in a general form that is able to account for the choice data, with one accumulator corresponding to a dim choice and the other to a bright choice that race independently. The model for the simple DRT is a special case with only one accumulator. For both simple and choice models, we add the possibility of variability in the starting point of evidence accumulation (see Logan, Van Zandt, Verbruggen, & Wagenmakers, 2014, for the mathematically equivalent case where the variation is in the threshold).

Evidence-Accumulation Modeling

A major division among evidence-accumulation models is whether they assume that accumulation is stochastic within a trial (i.e., the amount added to the evidence total in each moment during accumulation has a random component) or deterministic (i.e., the amount added in each moment is a constant). We fit a model of each type to the choice data, the deterministic LBA (Brown & Heathcote, 2008) and the stochastic racing one-barrier diffusion (Leite & Ratcliff, 2010) or Wald model, as illustrated in Figure 7. Our initial fit of the Wald race model assumed that the starting point of evidence accumulation varies from trial-to-trial according to a uniform distribution, an assumption shared with the LBA. The LBA differs from our Wald model in that the rate of evidence accumulation is assumed to vary randomly from trial to trial rather than from moment-to-moment within a trial. By fitting both models, we verified that the results are not dependent on specific assumptions of either models. As the LBA fits produced essentially the same conclusions we report details for the LBA in supplementary materials.

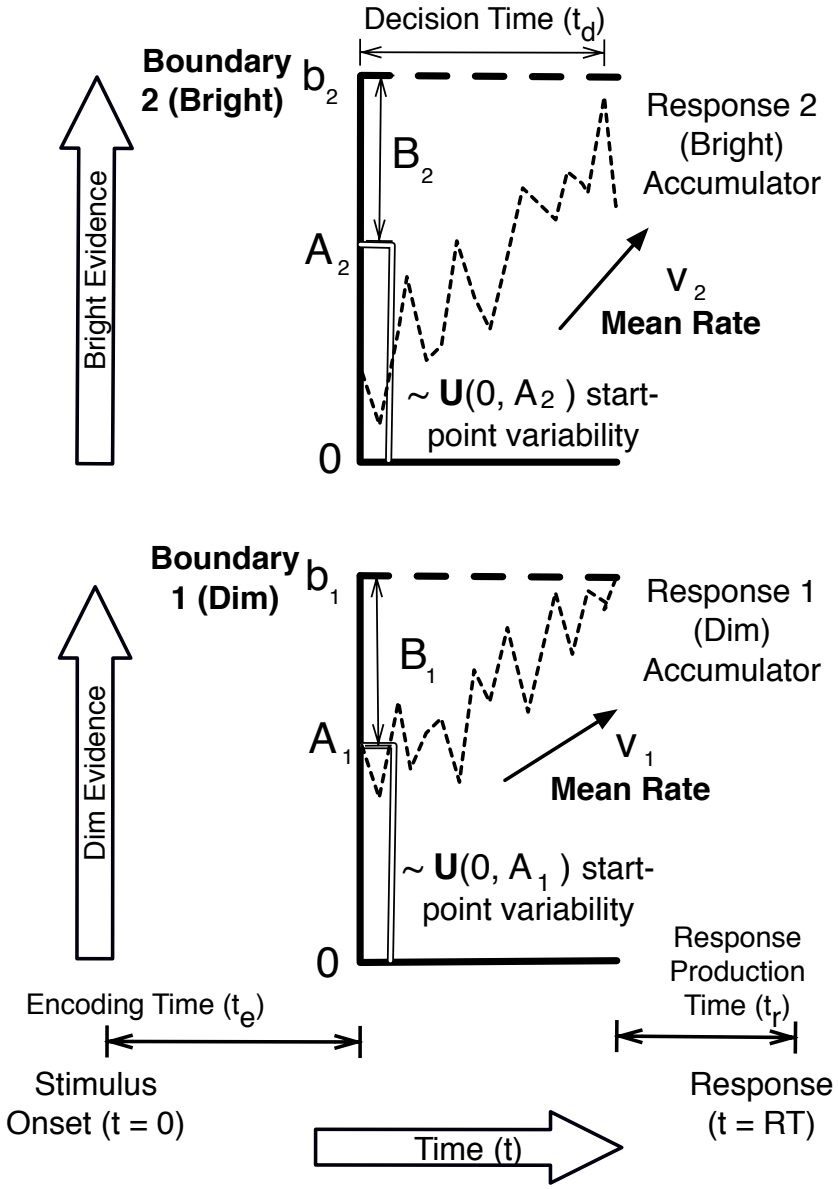

Figure 7. The Wald race model for the choice DRT with a dim stimulus and hence a higher rate for the matching (dim) accumulator than the mismatching (bright) accumulator. Note the dashed evidence accumulation paths are a caricature and would in reality vary more rapidly.

In order to identify the Wald model, we fixed the diffusion coefficient (i.e., the standard deviation of the moment-to-moment variability) at 1 and estimated the average rate of accumulation (*v*). One mean rate was estimated for the accumulator that matched the stimulus and a second, typically smaller, mean rate was estimated for the mismatching accumulator. For example, Figure 7 illustrates a case where the stimulus was dim and so the dim accumulator has

a higher mean rate than the bright accumulator. The starting point of evidence accumulation was assumed to vary from trial to trial independently for each accumulator according to a uniform distribution on the interval from 0 to an estimated parameter A. We parameterized each accumulator's threshold (b) in terms of the gap from the top of the start-point distribution, $B = b - A > 0$, so accumulation always began above the level of the start-point noise. We estimated non-decision time, the sum of the times to encode the stimulus and to produce a response, as parameter t_0.

Participants sometimes failed to respond to the DRT, particularly under increased load. One possibility is that this occurred because their response was so slow that it did not occur before the next DRT stimulus. However, for all participants in all conditions we found the right tail of their RT distributions clearly terminated well before the minimum interval between DRT stimuli of 3 seconds. This makes it very unlikely that the omissions were due to an ongoing decision process being cut off, at least if that decision process is of the same type as the one that explains all of the responses that were made. Therefore, we conceptualized response omissions as either a perceptual failure to encode the stimulus, akin to inattention blindness during distracted driving (Strayer & Drews, 2007), or as a failure to sample evidence from the encoded stimulus akin to the idea of "trigger failure" in models of the stop-signal paradigm (Matzke, Love & Heathcote, 2017, Matzke et al., 2017).

In particular, we adopted the standard evidence-accumulation model assumption that, at the onset of the stimulus, information transduction occurs through sensory processes that result in an internal representation of the relevant evidence (Brown & Heathcote, 2008; Ratcliff & McKoon, 2008). Then, the evidence accumulation process accesses this evidence. In the case of detection, this evidence consists of a simple change of sufficient magnitude in the attended sensory channel (Ratcliff & Van Dongen, 2011). Encoding failure means that either the stimulus

did not cause a change of sufficient magnitude, or that it did, but the change was not accessed by the evidence accumulation process. In order to account for the probability of these failures we augmented the model with a parameter, p_f. The omission probability can be directly observed and was clearly different between load conditions, and so p_f was assumed to vary with load in both models. Denoting the likelihood of a response R at time t in the standard models with parameter vector θ as $l(R,t|\theta)$, the corresponding likelihood in the augmented model is $(1-p_f)$ X $l(R,t|\theta)$, and the probability of an omission is p_f.

In the following sections we report a series of analyses based on fits of this model. We first report how several parameterizations of the model were fit to the simple DRT data and to the choice DRT data. Tables 4 and 5 report the results of analyses that enable selection of the best parameterization for each task type. We then report analyses of the parameters of the selected models, and of follow-up model selection and parameter analyses that allow us to identify the relative influence of each type of model parameter in explaining load effects. Finally, we look at the relationship between model parameters and tracking error.

Model Estimation and Selection

Model estimation was carried out in a Bayesian manner using the Differential Evolution algorithm (Turner, Sederberg, Brown, & Steyvers, 2013). Priors and sampling methods are described in Supplementary Materials. Sampling occurred in two steps. In the first step, sampling was carried out separately for individual participants. The results of this step provided the starting points for sampling the full hierarchical model, whose results are reported here (see Heathcote, Lin, Reynolds, Strickland, Gretton & Matzke, in press). Because we were primarily interested in the effects of cognitive workload, for both simple and choice DRTs we estimated separate threshold (B), mean rate (v), non-decision time (t_0) and omission probability (p_f) parameters for load-present and load-absent conditions, requiring a total of 8 parameters. We fit

the data for the two simple DRTs simultaneously, assuming the same non-decision time and omission probability parameters, but allowing different mean accumulation rates for the dim and bright stimuli for a total of four estimated mean rate parameters (dim and bright stimuli in load-present and load-absent conditions), and different boundaries, for a total of four estimated threshold parameters (i.e., dim and bright accumulators in load-present and load-absent conditions). For the choice DRT there were also four threshold parameters (allowing for response bias through different dim and bright accumulator boundaries in each of the load-present and load-absent conditions) but eight mean rate parameters, four for the matching accumulator (for dim and bright stimuli in load-present and load-absent conditions) and a corresponding four parameters for the mismatching accumulator.

We also compared three models that differed on whether start point noise (see Figure 7) was assumed to either be absent (i.e., $A = 0$), the same for all conditions and accumulators, or the same for all conditions but different between accumulators. Start point variability was selected for the simple DRT task, but not for the choice task. This outcome suggest that in the simple task participants prematurely sample evidence before the light appears, whereas the choice task they only sample evidence discriminating the choice after they detect the onset of the light (see Supplementary Materials for further discussion).

Comparison of 14 Wald Simple Detection Models

We refit the selected model (i.e., with start-point noise and with load effects on the probability of omission, rates, thresholds and non-decision time) and 6 variants that dropped one or two effects of load (except p_f, which always varied with load) with the lower bound on non-decision time set to 0.05s, but otherwise used the same priors as the initial fits. We did this also with the set same set of 7 models except we fixed start-point nose at zero. We used the same conventions to designate models: the most complex model, Bvt_0, has 13 parameters with start-

point noise and 12 without. With start-point noise the three further models dropped the load effect from one parameter, vt_0, Bt_0, and Bv, had 11, 11, and 12 parameters, respectively and one less each without start-point noise. Similarly, the final three models had a load effect on only one parameter, B, v, and t_0, had 10, 10 and 9 parameters respectively and one less each without start-point noise.

Table 4. The difference between DIC and the DIC for the best (Bvt_0 with start-point noise) model (DIC = -3599) and corresponding model weights for the set of 14 models.

		Bvt_0	Bv	Bt_0	vt_0	B	v	t_0
Start-point noise	DIC difference	0	193	316	483	855	957	2553
	Model Weight	.9988	0	0	0	0	0	0
No start-point noise	DIC difference	14	194	244	507	866	965	2578
	Model Weight	.0012	0	0	0	0	0	0

As shown in the Supplementary Materials, our results were consistent with the parameter analyses reported below in confirming load effects on all three parameters to be reliable (i.e., all models that dropped one or more effect were worse), both with and without start-point noise. The DIC for the selected model was very similar to the initial fit of this model, and analyses of parameter replicated a very similar pattern of p values.

Comparison of 7 Wald Choice Models

We again refit the selected model (i.e., with no start-point noise but with load effects on the probability of omission, rates, thresholds and non-decision time) and 6 variants that dropped one or two effects of load (except p_f, which always varied with load). To check the robustness and generality of our initial results all refits reduced the lower bound on non-decision time to 0.05s, but otherwise used the same priors as the initial fits. We denote models by parameters that vary with load, so the most complex model is Bvt_0. Three further models dropped the load effect from one parameter: vt_0, Bt_0, and Bv, with 14, 12, and 15 parameters respectively. The final three

models had a load effect on only one parameter, B, v, and t_0, with 11, 13 and 10 parameters respectively.

As shown in the Supplementary Materials, our results were consistent with the parameter analyses reported below in confirming load effects on all three parameters to be reliable (i.e., all models that dropped one or more effect were worse than the most complex Bvt_0 model). The sizes of the reductions in DIC suggest the non-decision time effect was least important, with the rate and threshold effects being equally important. The DIC for the model with all three effects was very similar to the initial fit of this model, and analyses of parameter replicated a very similar pattern of p values.

Table 5. The difference between DIC and the DIC for the best (Bvt_0) model (DIC = 10483) and corresponding model weights for the set of seven models.

	Bvt_0	Bv	Bt_0	vt_0	B	v	t_0
DIC difference	0	14	23	23	45	69	162
Model Weight	0.9991	0.0009	0	0	0	0	0

Parameter Tests

We report results about parameter estimates as posterior medians with 95% credible intervals given in square brackets and focus on the effects of load using Bayesian p-values to test differences in parameters between conditions (e.g., Matzke, Dolan, Batchelder &Wagenmakers, 2015; Matzke et al., 2017; Klauer, 2010; see supplementary materials for computational details). This p value is directly interpretable as the probability that one parameter is greater than another for the sample of subjects, so that a difference can be indicated by small or large p. However, given the familiar convention of low p values supporting a difference, we report the tail area such that small values are consistent with the stated effect direction.

Simple DRT. The response omission parameter (p_f) was 1.1%, [.37, 1.79] higher (6.1%, [5.6, 6.6] vs. 5.0%, [4.51, 5.50], $p = .002$), and non-decision time (t_0) was 0.023 s [.013, .032] faster (0.152 s, [.145, .158] vs. 0.175 s [.168, .180], $p < .001$) in the load-present condition than in the load-absent condition. The response threshold (b) was higher in the load-present than load-absent condition for both bright blocks (by 0.31, [.25, .36]: 1.13, [1.08, 1.18] vs. 0.83, [.78, .87], $p < .001$) and dim blocks (by 0.39, [.33, .44]: 1.20, [1.15, 1.24] vs. 0.81, [.77, .86], $p < .001$). The mean rate was clearly lower in the load-present than the load-absent condition for both bright blocks [.43, .69] (3.32, [3.21, 3.44] vs. 2.80, [2.68, 2.85], $p < .001$) and dim blocks [.06, .29] (2.91, [2.81, 2.99] vs. 2.72, [2.64, 2.81], $p < .001$).

Choice DRT. The proportion of response omissions was 2.1%, [1.41, 2.81] higher (3.4%, [2.90, 4.00] vs. 1.3%, [1.03, 1.71], $p < .001$), non-decision time was 0.031 s [.012, .051] faster (.218 s, [.201, .227] vs. 0.249 s, [.234, .263], $p < .001$), and the average response threshold was 0.27 [.200, .344] higher (2.1, [2.06, 4.15] vs. 1.83, [1.77, 1.89], $p < .001$) in the load-present condition than in the load-absent condition. The mean rate for the matching accumulator was clearly lower [.06, .24] in the load-present condition (2.32, [2.26, 2.38] vs. 2.46, [2.40, 2.53], $p < .001$), whereas the mismatching rate was a little higher [.003, .21] in the load-present condition (1.34, [1.27, 1.40] vs. 1.23, [1.16, 1.31], $p = .02$), so the difference between match and mismatch rates was much smaller [.15, .35] for the load-present condition than the load-absent condition (0.98, [.91, 1.05] vs. 1.23, [1.16, 1.30], $p < .001$).

Simple vs. Choice. We also compared the size of the selected models' load effects in simple and choice DRTs. There was no support for a decrease in the non-decision time load effect for simple compared to choice DRT ($M = .008$ s, $p = .23$, [-.03, .013]). However, there was some support for a larger threshold effect in the simple DRT ($M = .073$, $p = .05$, [-.015, .16]) and strong support for a larger rate effect ($M = .226$, $p = .001$, [.096, .35]).

The Underlying Causes of Load Effects

We used model selection to further test the necessity of rate, threshold and non-decision time effects in accounting for performance in both the simple and choice DRT (see supplementary materials for details). For the choice DRT we fit six simplifications of the selected models that removed the load effect on one or more parameters, and we compared the models using DIC. The analyses confirmed the results of the Bayesian *p*-value analyses, selecting the model allowing for load effects on accumulation rates, thresholds and non-decision time. We repeated this exercise for the simple DRT, using models both with and without start-point variability, and both confirmed the need for all three causes of the load effect and for the need for start-point noise.

We performed further analyses on the selected models in order to provide a more fine-grained quantitative understanding of the importance of each parameter in explaining the effects of load on speed, and for the choice DRT on accuracy. This is fairly straightforward for non-decision time, because it exclusively effects mean RT. In both simple and choice DRT the increased RT under load due to both higher thresholds and lower accumulation rates was masked somewhat by non-decision time, which decreased under load. In the simple DRT, it reduced the underlying load effect (i.e., the effect due to rate and threshold differences between load conditions) of 0.195 s by around 24% to the observed 0.148 s value. In the choice DRT it reduced the underlying 0.141 s mean RT load effect by a similar proportion, 22%, to the observed value of 0.11 s.

In order to quantify the effects of rate and threshold differences we modified the posterior parameter estimates from the selected models in two ways. First, we set the threshold parameters for the load-present and load-absent conditions to the same value, the average of the freely estimated parameters in the selected models. We then simulated data from this model and

calculated the predicted load effects in mean RT and, for the choice data, in accuracy, enabling us to quantify the effect of the remaining rate differences between load conditions. Second, we did the same but with rate parameters for load-present and load-absent conditions being set to their mean. Of the underlying 0.195 s simple DRT mean RT effect, around 0.12 s (60%) was due to the higher threshold in the load-present condition and the remaining 0.075 s to mean rate differences.

For the choice DRT, approximately the same length time, 0.12 s, but a larger proportion (85%) of the underlying 0.141 s underlying effect on mean RT was due to the threshold difference, with the remaining 0.021 s due to differences in mean rate. For the choice DRT, rate differences had a large effect on accuracy, increasing it in the load-absent over the load-present condition by 5.65%. However, this difference in accuracy was reduced 1.9% by the increase in threshold in the load-present condition, producing the observed value of 3.75% greater accuracy in the load-absent than the load-present condition.

Model Parameter and Pursuit Tracking Correlations

In order to relate steering performance and model parameter differences, we utilized plausible-value correlations (Ly et al., 2017). These are a fully Bayesian way to test correlations between subject-level covariates and hierarchical parameter estimates (the latter being the "plausible values"). The parameters came from the selected Wald models of both the choice and simple DRT, and the subject covariates are RMSE steering error for each participant. This analysis provides two sorts of estimates, and corresponding inferential procedures, assuming either a fixed-effects approach with inference specific to the sample of participants or a random-effects approach appropriate for generalizing inference to a new sample of participants. The latter approach provides a much more stringent test and so we focus on it, with details of all results provided in supplementary materials.

We analyzed correlations between RMSE and the posterior distributions of threshold (B), rate (v), omission rate (p_f), and non-decision time (t_0) separately in the load-present and load-absent conditions. For the choice model, we performed correlations with B averaged across the dim and bright light accumulators, with both the difference between matching and mismatching accumulator rates (as a measure of the quality of evidence), and with the rate for the matching accumulator (which is most strongly related to the speed of correct responses and most directly analogous to the simple DRT rate). As for the previous modeling methods, 95% credible intervals are given in square brackets. Bayesian p-values are based on the distribution of posterior parameter estimates of correlations with RMSE. In order to maintain the convention that smaller p-values support the existence of differences, we give the probability of a correlation being greater than zero for negative correlations, and the probability of a correlation being less than zero for positive correlations; A p near zero value supports there being a strong negative or positive relationship and a p value near .5 supports there being no relationship.

For the choice data there was a clear negative correlation between steering error and the rate for the matching accumulator in the load-present condition, $r(17) = -.61$, [-.86, -.24], $p = .003$. The analogous correlation for the load-absent condition was moderately large, but its 95% credible included zero, $r(17) = -.40$, [-.75, .05], $p = .041$. The same was true for a negative correlation with the difference between match and mismatch rates in the load-present condition, $r(17) = -.41$, [.78, .02], $p = .031$ and for a positive correlation with the omission probability (p_f) parameter in the load-present condition, $r(17) = .45$, [-.01, .77], $p = .028$. Correlations with non-decision times and threshold were generally weak.

For the simple DRT a negative correlation with mean rate was present with the dim stimulus for the load-present condition $r(17) = -.46$, [-.78, -.03], $p = .021$, analogous to the results in the choice DRT. This correlation was weaker for the bright (i.e., ISO standard) DRT,

$r(17) = -.38$, [-.73, .07], $p = .051$. The same was true for those correlations in the load-absent conditions for both the dim DRT, $r(17) = -.42$, [-.76, .03], $p = .036$ and the bright DRT, $r(17) = -.41$, [-.75, .04], $p = .039$. For the bright DRT the strongest finding was a negative correlation with the non-decision time in the load-absent condition, $r(17) = -.45$, [-.77, -.01], $p = .024$. The analogous correlation was of similar magnitude but a little weaker for the dim DRT, $r(17) = -.41$, [-.75, .03], $p = .035$, as were the correlations with non-decision time in the load-present condition for both the bright $r(17) = -.36$, [-.72, .09], $p = .057$, and dim $r(17) = -.38$, [-.73, .07], $p = .051$ DRTs. For both bright and dim DRTs correlations with thresholds and omission probability were weak.

General Discussion

It is a fundamental characteristic of human cognition that dividing attention between two or more tasks results in performance decrements (i.e., slower and more error-prone behavior) compared to when each task is performed separately. The International Standards Organization developed the DRT (ISO DIS 17488, 2015) to assesses cognitive workload in a variety of multitasking situations. DRT reaction time and omission rates are very sensitive to increases in cognitive workload; however, the precise reason is unclear. This sensitivity could be due to cognitive-capacity related changes in the rate of evidence accumulation, or to a strategic adjustment in the threshold amount of information required to trigger a response, changes in non-decision time (i.e., the time to encode a stimulus and to produce a response), or some combination of these factors. These distinctions could meaningfully change the approaches to applications of studying cognitive workload, such as alleviating driver distraction. For example, approaches that target attention allocation from a limited-resource perspective would be validated with a demonstration of rate effects. Threshold effects may call into question current assumptions about cognitive workload, and subsequently shift focus toward individual

differences in strategic decision making. Increased non-decision time effects would imply cognitive workload is mainly due to early processing or subsequent motor interference. The current research used formal modeling to identify the bases for changes in DRT performance with increased cognitive workload.

We found that in both choice and simple DRTs the cognitive workload induced by a secondary task of counting backward by threes (while also performing a primary steering task) reduced evidence accumulation rates. These results suggest that information processing in choice and simple DRTs depends on the same limited pool of attention capacity as the secondary task. To our knowledge, this is the first direct confirmation that cognitive workload, as traditionally measured by a dual-task methodology, affects evidence accumulation rates. This finding was bolstered by its consistency in two modeling frameworks: the shifted Wald (Heathcote, 2004, Logan et al., 2014), and the LBA (Brown & Heathcote, 2008, see supplementary materials), and by its consistency between the ISO DRT using a bright light and two variants: either requiring detection of a dim light or requiring a binary choice between bright and dim lights. It confirms Strayer et al.'s (2011, 2015) interpretation of correlations between the DRT and effects of secondary tasks on driving performance as being at least in part mediated by limited-capacity attention.

These results support the dominant assumption that cognitive workload effects reflect a competition for limited resources. Further supporting the notion of capacity sharing, the choice DRT clearly reduced performance in the pursuit-tracking task. This was also true to a lesser degree for simple DRT using a dim light, with the smallest, but still reliable, impact being produced by the ISO DRT using a bright light. As well as confirming the notion that the DRT draws on the same limited-capacity attention pool as the primary and secondary tasks, it also

confirms that the ISO DRT with an easily detected stimulus is best suited to minimize the impact of measuring cognitive workload during driving.

Increased cognitive workload also causes an increase in DRT response omissions, that is, failures to respond to the present DRT stimulus before the onset of the next DRT stimulus. It was clear in our data that omissions were not simply due to slow responses coming from the same process producing observed DRT responses, as the distribution of observed DRT responses terminated well before the next stimulus appeared. We accounted for response omissions by assuming a mixture of normal Wald evidence accumulation processes and failures to encode the DRT stimulus. Our results imply that this encoding failure process is sensitive to cognitive workload.

Ratcliff and Strayer (2014) took a different approach to omissions, using a Wald model, but assuming Gaussian trial-to-trial variability that sometimes results in negative accumulation rates and hence response omissions, because the accumulated evidence cannot reach the positive threshold. However, this came at a cost. The model has no closed-form likelihood, so had to be fit by slow simulation-based methods. More importantly, it has problems with parameter identification in the simple DRT, meaning it cannot adjudicate whether a threshold effect, a rate effect, or both, mediate slowing. We demonstrated our model does not suffer from the same problems. In supplementary materials we report extensive parameter-recovery (Heathcote et al., 2015) simulations, which show that the Wald model produces quite accurate and precise estimates of parameters relevant to cognitive workload effects with samples as small as 200 trials per participant. This makes our model practical to apply to an ISO DRT recorded over a duration as short as 15 minutes. These outcomes increase the feasibility of applying evidence-accumulation modeling techniques to cognitive workload measurement in a wide range of behavioral tasks for both laboratory and applied settings.

Contrary to our initial hypothesis, we found a small but reliable *decrease* in non-decision time under cognitive workload. This may have been compensatory in nature, slightly offsetting (by about ~20%) slowing due to threshold increases and rate decreases. Palada, Neal, Tay and Heathcote (2018) also found that high cognitive workloads could sometimes be associated with reduced non-decision time in a difficult choice task. They suggested that fast non-decision times were associated with a degraded encoding. It is possible that such a degraded encoding process might in part be responsible for the reduced rate of evidence accumulation we observed under cognitive workload (as a weakened stimulus encoding weakens the evidence on which the rate is based). However, for Palada et al. the decrease in non-decision time was more extreme, it was associated with a drastic decrease in accuracy specific to one type of choice response, and only occurred under extreme time pressure that caused participants to run out of time to respond to some of the multiple stimuli they had to respond to in each display. Hence, further research is required to determine if the same mechanism is in play in the very different task setup used here.

We found a clear increase in both simple and choice DRT thresholds due to a secondary-task workload. Tillman et al. (2017) also found conversation on a hands-free cell-phone caused an increase in threshold in the Wald model of the ISO DRT. However, they did not find any effect of this secondary task on accumulation rates or non-decision time, despite its well-documented deleterious effects on a primary driving task. They suggested their DRT slowing and threshold increase was an indirect result of a general tendency to be more cautious when making responses in more demanding situations. It seems likely that the same mechanism may have been at play in our results. However, in contrast to Tillman et al., we found that a little less than half of the slowing due to load was due to a decreased rate of evidence accumulation in the simple DRT. In contrast, our results in the choice DRT were more in line with theirs, with threshold

effects explaining the majority (~85%) of the slowing, although even in this case we obtained clear evidence for a reliable rate effect.

Once again, these results support the use of a simple DRT as best for measuring cognitive workload, as it better reflects accumulation rate effects that indicate a decrease in available attention capacity that could affect driving performance. At a theoretical level, the divergence between our results and those of Tillman et al. (2017) clearly indicates slowing in the DRT is not by itself sufficient to make inferences about underlying causes. Fortunately, our model provides a practical and efficient way to make such inferences in future research. We provide the software necessary to do so through the Open Science Framework (https://osf.io/e8kag/).

Correlations between parameter distributions and pursuit tracking error provided further evidence of a relationship between accumulation rates and driving performance. Accumulation rates in choice and simple DRTs correlated negatively with steering performance, whereas we did not find evidence to support a correlation between thresholds and steering performance. These results demonstrate a relationship between individuals with less cognitive capacity (as reflected in lower DRT accumulation rates) having higher steering errors, particularly under more demanding high-load conditions and with the more demanding dim and choice DRTs. The weaker correlations for the less demanding ISO DRT reinforce our finding that it has a lesser impact on the steering task.

In summary, the DRT has been shown to be very sensitive to dynamic changes in cognitive workload in a variety of multitasking contexts (e.g., Strayer, Biondi, & Cooper, 2017). Here, we have provided a theoretical account for what aspects of information processing are captured by the technique. The DRT is particularly slowed with cognitive workload due to a decrease in the rate of evidence accumulation, an effect that is substantially accentuated by an increase in the amount of evidence required to trigger a response but somewhat masked by a

decrease in non-decision time. In terms of applications to distracted driving, these findings support strategies and policies that optimize a driver's allocation of limited resources to the road. However, they also suggest there is scope for improving driving performance by encouraging compensatory strategies that give non-driving tasks lower priorities through changing the amount of information required to make the associated choices and/or by making it easier to encode information for the non-driving tasks, and through training to decrease motor production times for the associated responses.

In closing, this research provides a framework for accurately quantifying cognitive workload and the factors that contribute to it, which will allow future researchers and policy makers to determine the danger inherent in many tasks within the vehicle. Additionally, it is possible that future experimental manipulations will alter the information processing dynamics underlying the DRT. The formal modeling described herein will assist in identifying changes to the factors underlying cognitive workload and allow this framework to be flexibly applied across multiple paradigms.

References

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*(1), 158-175.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.

Biernat, M., Kobrynowicz, D., & Weber, D. L. (2003). Stereotypes and shifting standards: Some paradoxical effects of cognitive load. *Journal of Applied Social Psychology*, *33*(10), 2060-2079.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153-178.

Caird, J. K., Willness, C. R., Steel, P., & Scialfa, C. (2008). A meta-analysis of the effects of cell phones on driver performance. *Accident Analysis & Prevention*, *40*(4), 1282-1293.

Castro, S., Cooper, J., & Strayer, D. (2016, September). Validating two assessment strategies for visual and cognitive load in a simulated driving task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, No. 1, pp. 1899-1903). Sage CA: Los Angeles, CA: SAGE Publications.

Castro, S. (2017, September). How Handheld Mobile Device Size and Hand Location May Affect Divided Attention. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 1370-1374). Sage CA: Los Angeles, CA: SAGE Publications.

Cooper, J. M., Castro, S. C., & Strayer, D. L. (2016, September). Extending the detection response task to simultaneously measure cognitive and visual task demands. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, No. 1, pp. 1962-1966). Sage CA: Los Angeles, CA: SAGE Publications.

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,* 42-45.

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review*, *16*(6), 1129–1135.

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010). Converging measures of workload capacity. *Psychonomic Bulletin & Review*, *17*(6), 763–771.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology, 52*, 139-183.

Heathcote, A. (2004). Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S-PLUS package. *Behavior Research Methods*, *36*, 678–694.

Heathcote, A., Brown, S.D. & Wagenmakers, E-J. (2015). An introduction to good practices in cognitive modeling. In Forstmann, B. U., & Wagenmakers, E.-J. (Eds.). *An Introduction to Model-Based Cognitive Neuroscience*. Springer, New York.

Heathcote, A., Lin, Y-S, Reynolds, A., Strickland, L., Gretton, M. & Matzke, D. (in press). Dynamic models of choice. *Behavior Research Methods*.

Heathcote, A., Loft, S., & Remington, R. W. (2015). Slow down and remember to remember! A delay theory of prospective memory costs. *Psychological Review, 122*(2), 376-410.

ISO DIS 17488 (2015). Road Vehicles -Transport information and control systems - Detection Response Task (DRT) for assessing selective attention in driving. Draft International Standard, ISO TC 22/SC39/WG8.

Kahneman, D. (1973). *Attention and effort* (p. 246). Englewood Cliffs, NJ: Prentice-Hall.

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent–trait approach. *Psychometrika, 75* , 70–98.

Klauer, S. G., Guo, F., Simons-Morton, B. G., Ouimet, M. C., Lee, S. E., & Dingus, T. A. (2014). Distracted driving and risk of road crashes among novice and experienced drivers. *New England Journal of Medicine*, *370*(1), 54-59.

Laming, D.R.J. (1968). *Information theory of choice-reaction times.* London: Academic Press.

Leite, F. P., & Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, & Psychophysics*, *72*(1), 246–273.

Levy, J., Pashler, H., & Boer, E. (2006). Central interference in driving: Is there any stopping the psychological refractory period?. *Psychological Science*, *17*(3), 228-235.

Logan, G. D., Van Zandt, T., Verbruggen, F., & Wagenmakers, E. J. (2014). On the ability to inhibit thought and action: General and special theories of an act of control. *Psychological Review, 121*(1), 66-95. doi:10.1037/a0035230.

Ly, A., Boehm, U., Heathcote, A., Turner, B.M., Forstmann, B., Marsman, M., & Matzke, D. (2017). A flexible and efficient hierarchical Bayesian approach to the exploration of individual differences in cognitive-model-based neuroscience. In A.A. Moustafa (Ed.) *Computational models of brain and behavior* (pp. 467-480). Wiley Blackwell.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd edition). Lawrence Erlbaum Associates. New York.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from plausible values? *Psychometrika*, *81*(2), 274-289.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). Tutorial in *Quantitative Methods for Psychology, 4*, 61-64.

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*, 205–235.

Matzke, D., Hughes, M., Badcock, J.C., Michie, P. & Heathcote, A. (2017). Failures of cognitive control or attention? The case of stop-signal deficits in Schizophrenia, *Attention, Perception & Psychophysics*, *79*, 1078-1086.

Matzke, D., Love, J. & Heathcote, A. (2017). A Bayesian approach for estimating the probability of trigger failures in the stop-signal paradigm. *Behavior Research Methods*, 49, 267–281.

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, *126*(2), 247.

National Highway Transportation Safety Administration (2012). Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices. *Department of Transportation*, Docket No. NHTSA-2010-00053.

National Highway Traffic Safety Administration (2016, November 21). Statement, and Notice of Proposed Visual-Manual NHTSA Driver Distraction Guidelines for Portable and Aftermarket Devices, Docket No. NHTSA-2013-0137. Retrieved from https://www.federalregister.gov/documents/2016/12/05/2016-29051/visual-manual-nhtsa-driver-distraction-guidelines-for-portable-and-aftermarket-devices

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review, 86*(3), 214.

Palada, H., Neal, A., Tay, R., & Heathcote, A. (2018, submitted). Understanding the causes of adapting, and failing to adapt, to time pressure in a complex multi-stimulus environment.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Ranney, T. A., Baldwin, G. H., Smith, L. A., Mazzae, E. N., & Pierce, R. S. (2014). *Detection Response Task (DRT) Evaluation for Driver Distraction Measurement Application* (No. DOT HS 812 077).

Ratcliff, R. (2015). Modeling one-choice and two-choice driving tasks. *Attention, Perception, & Psychophysics*, *77*(6), 2134–2144.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., & Strayer, D.L., (2014). Modeling simple driving tasks with a one-boundary diffusion model. *Psychonomic Bulletin & Review*, *21*(3), 577-589. doi:10.3758/s13423-013-0541-x.

Ratcliff, R., & Van Dongen, H. P. A. (2011). Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(27), 11285–11290. http://doi.org/10.2307/27978781

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*(3), 414–429.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.s

Strayer, D. L., Biondi, F., & Cooper, J. M. (2017). Dynamic workload fluctuations in driver/non-driver conversational dyads. In D. V. McGehee, J. D. Lee, & M. Rizzo (Eds.) *Driving*

Assessment 2017: International Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design (pp. 362-367). Published by the Public Policy Center, University of Iowa.

Strayer, D. L., & Drews, F. A. (2007). Cell-phone–induced driver distraction. *Current Directions in Psychological Science*, *16*(3), 128-131.

Strayer, D. L., & Fisher, D. L. (2016). SPIDER: A model of driver distraction and situation awareness. *Human Factors, 58*, 5-12.

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J., Medeiros-Ward, N., & Biondi, F. (2015). Assessing cognitive distraction in the automobile. *Human Factors, 53*, 1300-1324.

Strayer, D. L., Watson, J. M., & Drews, F. A. (2011). Cognitive distraction while multitasking in the automobile. *Psychology of Learning and Motivation-Advances in Research and Theory*, *54*, 29-58. Academic Press.

Strickland, L., Loft, S., Remington, R.W. & Heathcote, A. (in press). Racing to remember: A theory of decision control in event-based prospective memory. *Psychological Review*.

Strickland, L., Loft, S., Remington, R.W. & Heathcote, A. (2017). Accumulating evidence for the delay theory of prospective memory costs, *Journal of Experimental Psychology: Learning, Memory & Cognition*, *43, 1616-1629*.

Tillman, G. Strayer, D., Eidels, A., Heathcote, A. (2017). Modeling cognitive load effects of conversation between a passenger and driver, *Attention, Perception & Psychophysics 79*(6), 1795-1803.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368–384.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.

Welford, A. T. (1952). The ‘psychological refractory period’ and the timing of high-speed performance—a review and a theory. *British Journal of Psychology*, *43*(1), 2-19.