Revision of MS 2017-0646 as invited by the action editor Pierre Barrouillet

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 including Abstract: 5958

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 detection and choice versions of the DRT in a driving scenario. Our experiments also 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 the most 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. The data provide converging evidence that the DRT method is sensitive to dynamic fluctuations in limited capacity attention.

Keywords: Detection Response Task, Driving Simulation, Linear Ballistic Accumulator, Independent Race Model, Cognitive Workload

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 divide their attention, they attempt to perform more than one mental activity at the same time (e.g., driving an automobile and using a smartphone to talk, text, etc.). Although robust, the precise cause of the capacity limitations 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. On the other hand, declines in performance may reflect a more conservative response *bias* under higher cognitive workload. Though 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 *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.

In fact, 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 deficits in attention—of which the cognitive workload imposed by mobile device use is a major cause—are a leading factor in the majority of crashes and near-crashes. 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.

Kahneman (1973) described 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. 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 directly correlated 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. With appropriate apparatus (i.e., stimuli and responses that do not overlap with other tasks) it has only minimal effects on driving. 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).

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., Heathcote, 2004; 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 the sum of encoding and response production times (non-decision time).

Given that cognitive workload is thought to affect the rate of information processing, it naturally maps to the rate parameters of these models. However, Heathcote, Loft, and Remington (2015) demonstrated in the domain of prospective memory—which was thought to slow present

tasks by sharing resources with a future goal—that slowing of task performance stemmed primarily from individuals delaying responses of ongoing tasks in order for the possibility of a prospective memory task response to compete with the more salient repetitive response. This conclusion calls into question prevailing theories of prospective memory and the effect of a limited capacity system. In the domain of cognitive workload, it is possible that workload slows early perceptual encoding, and hence non-decision time, or even causes failures to encode evidence from the stimulus, and hence response omissions. In two-choice tasks, the quality of evidence—which is both inversely proportional to the level of noise in evidence and directly proportional to the difference between evidence for the options—may also be affected by attention. This is true both in race models, where separate accumulators for each option trigger their corresponding response if they are first to reach threshold (Brown & Heathcote, 2008, Leite & Ratcliff, 2010) and in single accumulator models that directly accrue evidence differences (Ratcliff & Rouder, 1998). 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", Ratcliff & Rouder, 1998). Higher noise in evidence associated with reduced attention may also cause false detection responses, and so again participants may set larger thresholds with the increased cognitive workload. 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.

Visual load effects on accumulation rates in perceptual choice tasks were found by Eidels, Donkin, Brown and Heathcote (2010). Schmiedek et al. (2007) also conceptualized correlations between individual differences on a variety of tasks (working memory, reasoning, and psychometric speed) and evidence-accumulation rates in verbal, numeric and spatial choice

paradigms regarding attention capacity. However, we are aware of no previous studies that have directly assessed the relationship between cognitive workload and rate parameters, or indeed any other parameters of evidence-accumulation models, either in choice or simple detection tasks. 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 the primary task regarding perceptual and motor components. We performed such a dual-task experiment, with a driving-like primary task and a secondary task of counting backward by threes.

In a baseline condition no workload measurement was taken, and in a second condition, the 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 workload was measured with a choice version of the DRT, where from trial-to-trial participants had to press one of two buttons to indicate whether the light was dim or bright. We hypothesized that in both tasks the secondary-task load would slow DRT responding and increase errors (both choice errors and missed responses). From past work on visual workload and individual differences, we hypothesized that in both tasks cognitive workload would reduce evidence accumulation rates. More tentatively, we hypothesized that cognitive workload might 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. Because our models were fit to each distribution of responses per the conditions of each participant separately, and all

comparisons were completely within participants, the number of trials per condition was critical for model estimation, and not the number of participants (see supplementary materials for parameter recovery, suggesting adequate sample size and quality of data).

Materials

A 106 cm Samsung LCD (1920 x 1080 pixels) was used to display the tracking task. Participants utilized a steering wheel from a driving simulator to track a ball that moved continuously on the screen. The steering wheel updated the location of the cursor through a Sparkfun™ Electronics rotary encoder set to sample the position at 200Hz. The DRT device presented a dash-mounted light at two intensities of red. Stimuli were presented randomly every 3-5 seconds and responses were made by pressing one of two micro-switches attached to participant's left and right thumbs.

Design

The 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. The probability of the ball's location followed a normal distribution so that the ball moved smoothly through the center third of the screen approximately 68% of the time, and the center two thirds approximately 95% of the time.

There were four tracking conditions: single-task tracking, and tracking and concurrently making a detection response to the onset of a low-intensity light or a high-intensity light or a choice response to a low- vs. high-intensity light. The same stimulus-to-response mapping was used in both the choice task and the detection tasks. These conditions were crossed fully with backward counting by threes (3s) or not counting (None). The 4 x 2 factorial design was blocked into 64 one-minute runs and counterbalanced using a Latin Square. Participants were given 30

seconds of rest between each block. Apart from single-task tracking, there were an average of 240 DRT trials for each of the cells of the design.

Calibration. Before the experiment, the lights were calibrated for each participant so that they were approximately 75% accurate in their choice classification. 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 high and low-intensity lights to 200 and 100 respectively. Participants made sets of 8 choice responses; then the low-intensity 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. 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 intensities after that remaining fixed.

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 successful responses out of 15 were also removed (1.20%), as were blocks with lower than 50% accuracy, or proportion of hits to misses (0.76%, 0.09%, respectively). Root Mean Squared Error (RMSE) tracking error was computed from differences between the position of the cursor and the target sampled at 200Hz. The 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). We first report conventional analyses of tracking error, the proportion of missed responses, accuracy and mean RT using the

lme4 package (Bates, Maechler, Bolker, & Walker, 2015). Participants were included as a random effect, and effects were assessed using a Type II Wald chi-square test. We report 95% confidence intervals in square brackets.

Pursuit Tracking. As shown in Figure 1, RMSE steering error was greater when counting (M = 2.23, [2.21, 2.24]), than when not counting (M = 2.16, [2.14, 2.17]), $\chi^2(1)$ = 157.92, $p < .001$. Relative to the single-task condition (M = 1.97, [1.96, 1.99]) the ISO standard DRT (i.e., high) increased steering error (M = 2.15, [2.14, 2.17]), $t(16) = 3.83$, $p = .001$, [.08, .28], but had a smaller steering error effect than the dim DRT stimulus (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 low DRT stimulus as well, $t(16) = 3.65$, $p = .002$, [.06, .24]. The average load effect with the addition of the detection tasks did not differ significantly from the load effect in the simple task, $t(16) = 0.75$, $p = .46$, [-.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 $\chi^2(3) = 14.75$, $p = .002$ [.01, .25].

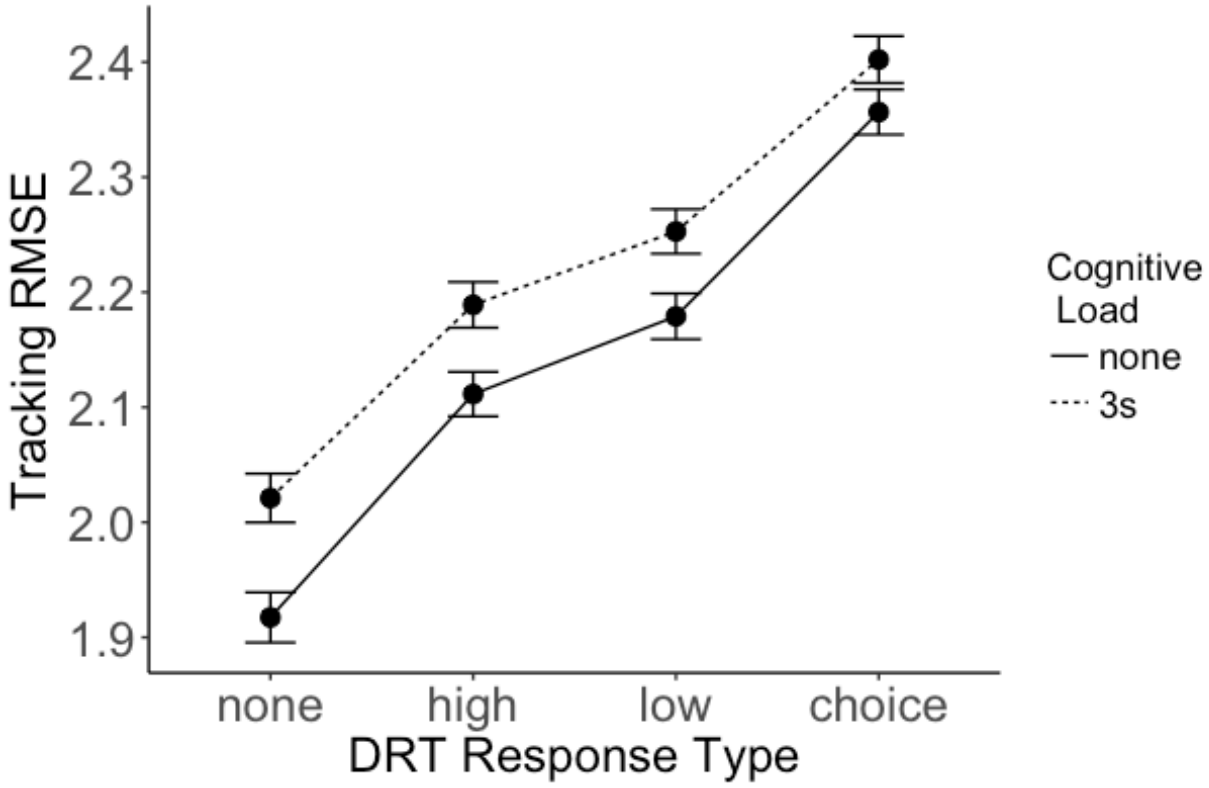

Figure 1. Root Mean Squared Error (RMSE) for the tracking task. Error bars are 95% confidence intervals around the mean, calculated as described in Morrey, (2008).

Misses. We combined data from the two detection DRTs to make 2 (stimulus) × 2 (load) design. Miss rates were lower when participants were not counting (M = 4.30%, [3.70, 4.89]) than when they were counting (M = 5.96%, [5.24, 6.67]), $\chi^2(1) = 18.66$, $p < .001$, but neither the effect of stimulus type, nor its interaction with load, was significant. For the same design, but replacing the simple conditions with choice, load also significantly affected miss rates, $\chi^2(1) = 36.64$, $p < .001$. Participants failed to respond more often when counting (M = 3.60%, [2.73, 4.47]) than when not counting (M = 1.31%, [.80, 1.83]), but neither the effect of stimulus type, nor the interaction with load, was significant. When the detection and choice data were combined the increase in misses due to load was significantly greater for the choice (2.00%, [1.57, 3.01]) than the detection (1.10%, [.96, 2.36]) task, $\chi^2(1) = 10.62$, $p = .001$.

Choice Accuracy. Participants were more accurate when not counting (M = 77.40%, [75.48, 79.27]) than when counting (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.

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 without transformation. Participants responded 0.146s slower when counting (M = 0.622s, [.614, .631]) than when not counting (M = 0.479s, [.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 0.018s slower for the low light stimulus (M = 0.555s, [.546, .561]) than the high light stimulus (M = 0.539s, [.529, .543]). The two effects did not interact significantly.

In the choice task participants responded 0.098s slower when counting (M = .966, [.951, .980]) than when not counting (M = .868s, [.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—0.019s slower for the low light stimulus (M = .923s, [.909, .937]) compared to the high light stimulus (M = .904s, [.891, .917])—and again the two effects did not interact significantly. We also found that participants were 0.049s slower overall on error (M = .951s, [.929, .973]) than correct trials (M = .902s, [.891, .912]), $\chi^2(1) = 34.04$, $p < .001$.

When the detection and choice data were combined detection was much quicker overall (M = .545s, [.540, .550] vs .913s, [.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 detection (M = .151s, [.140, .162]) than the choice *(*M = .102s, [.088, .116]) task, $\chi^2(1) = 212.13$, $p < .001$. The RMSE tracking error and RT exhibit the same *underadditive* pattern with an increase in cognitive workload and DRT difficulty. With the addition of the discrimination task for the choice DRT,

both responses to the DRT and RMSE tracking error are less affected by the cognitive workload manipulation.

Modeling Choice

We assumed two accumulators representing each choice option (high vs. low) that raced independently to determine the response. 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 processes (Leite & Ratcliff, 2010). We call the latter a "Wald" model because the distribution of times for an accumulator to reach threshold follows a Wald distribution. In particular, we assumed a shifted Wald (Heathcote, 2004, shows this model has good estimation properties) where the non-decision time is a constant. Following Logan, Van Zandt, Verbruggen and Wagenmakers (2014) our initial fit of the Wald model also assumed that the starting point of evidence accumulation varies from trial-to-trial according to a uniform distribution. Both assumptions are shared with the LBA.

Model Specification. In the LBA model, each accumulator had starting points uniformly distributed in the interval 0-A that was drawn independently for each accumulator on each trial. The same value of start-point noise (A) parameter was assumed for all conditions. Evidence accrues linearly and deterministically at a rate drawn, independently for each accumulator and trial, from a normal distribution truncated below by zero with mean v and rate standard deviation s_v. In order to identify the model (Donkin, Brown & Heathcote, 2009), the s_v parameter was fixed at 1 for the accumulator that mismatched the stimulus (i.e., the high accumulator when the

stimulus was low intensity and the low accumulator when the stimulus was high intensity), but s_v for the other (matching) accumulator was estimated. The *v* parameter was estimated for matching and mismatching accumulators for each stimulus (high and low). The threshold *(b*) was allowed to vary between accumulators to accommodate potential response biases (e.g., a lower threshold for the bright accumulator would cause a bias to respond "bright"), and parameterized concerning the gap (*B*) between the top of the start point noise and threshold (i.e., $B = b - A$), with *B* allowed to vary with load. Non-decision time (t_0) was assumed the same for both accumulators, but allowed to vary with load.

Following Logan et al. (2014), 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. The Wald model made the same assumptions about non-decision time (t_0), starting-point distributions (*A*), evidence-accumulation rates (*v*), and thresholds (*B*) as the LBA. Note that the LBA and Wald models each assume two sources of noise, one held in common (start-point noise) and one not: moment-to-moment rate noise in the Wald and trial-to-trial rate noise in the LBA.

We augmented both models with a parameter, p_f, to account for response omissions (i.e., the probability of failing to respond on the DRT). The omission rate 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 $(1-p_f) \times l(R,t|\theta)$ and the probability of an omission is p_f.

Modeling Methods. Estimation using the likelihood just defined 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. The

models provided an accurate description of the data; details of goodness-of-fit are also provided in supplementary materials.

We provide 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., 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 in a way such that small values are consistent with the stated effect direction.

We compared different models using the Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). Smaller values indicate a better model regarding providing the best tradeoff between simplicity and goodness of fit. DIC is on a logarithmic scale, so a difference of 10 can be considered strong evidence in favor of a model, and DIC values can be converted to model weights (*w*, Wagenmakers & Farrell 2004), which under the appropriate prior and assuming the true model is in the set of models under consideration approximate the probability that a model in a given set of models is the data generating model.

LBA Results. The response omission parameter was higher in the 3s condition by 2.1%, [1.5, 2.8] (3.4%, [2.9, 4.0] vs. 1.3%, [1.0, 1.7], $p < .001$), non-decision time was 0.014s [.002, .028] faster (.130s, [.122, .141] vs. 0.144s, [.132, .156], $p = .042$), and the response threshold was 0.15, [.08, .22] higher (averaged over high and low accumulators, 1.36, [1.30, 1.46] vs. 1.21, [1.15, 1.26], $p < .001$). The average rate for the matching accumulator, which largely determines

the speed of correct responses, was clearly lower [.12, .26] in the 3s condition (2.3, [2.24, 2.37] vs. 2.5, [2.43, 2.56], $p < .001$), whereas the mismatching rate was only marginally lower [-.05, .15] in the 3s condition (1.24, [1.16, 1.33] vs. 1.29, [1.22, 1.37], $p = .17$). The difference between match and mismatch rates, which largely determines accuracy, was smaller [.05, .24] for the 3s condition (1.06, [.99, 1.13] vs. 1.2, [1.14, 1.27], $p < .001$).

Overall, these results indicate that more frequent choice errors and slower responding in the 3s condition was due to a reduced accumulation rate. The tendency for increased choice errors was partially counteracted by a higher threshold, which also exaggerated the slowing in the 3s condition, consistent with participants attempting to trade speed for accuracy to counteract the load effect. The slowing under load was also counteracted by a decrease in non-decision time, but this effect was small; without it slowing would have been 0.127s rather than 0.113s in median RT, a decrease of only 11%.

Wald Results. Model selection clearly favored the Wald (DIC =10512) over the LBA (DIC = 10564), $w > .999$. Because the idea of adding start-point noise to the Wald model is relatively new, having been first suggested by Logan et al., (2014) in the context of a much more complex model of performance in the stop signal paradigm, we fit Wald models both with no start-point noise, and with start-point noise varying between the two choice accumulators. The latter model has an improved DIC (10500), but the model with no start-point noise was the winner (DIC = 10488). Given this clearly better performance ($w = 0.996$ among the three Wald models), we focused on the 16-parameter model with no start-point variability. Despite having 2 less parameters per participant, this model fits at least as well as the LBA model (see supplementary materials).

Results of parameter analyses for the selected Wald model revealed a pattern of effects very similar to the LBA. The proportion of response omissions was 2.1%, [1.41, 2.81] higher in the 3s condition (3.4%, [2.90, 4.00] vs. 1.3%, [1.03, 1.71], $p < .001$), non-decision time was 0.031s, [.012, .051] faster (.218s, [.201, .227] vs. 0.249s, [.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], $p < .001$). The average rate for the matching accumulator was clearly lower [.06, .24] in the 3s 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 3s 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 3s condition (0.98, [.91, 1.05] vs. 1.23, [1.16, 1.30], $p < .001$).

Thus both the LBA and Wald models confirmed that increased RT and decreased choice accuracy under load was primarily due to a rate difference, with a speed-accuracy tradeoff due to a lower threshold somewhat increasing the RT effect and decreasing the accuracy effect. The decrease in the observed load effect on median RT due to speeding in non-decision time was estimated to be a bigger than under the LBA model, but was still not large at only 22%.

To further test the roles of rate, threshold and non-decision time effects we fit six simplifications of the selected model that removed the load effect on one or more parameters. This confirmed the results of the Bayesian *p*-values, selecting the model allowing for load effects on rates, thresholds and non-decision time (see supplementary materials for details), but indicating that the latter effect was least important, with the rate and threshold effects equally important.

Modeling Detection

We fit the simple detection data using the same method of accounting for response omissions as with the choice data, but assuming a model with only one accumulator. Initial analyses indicated that, that a Wald accumulator model was clearly preferred over an LBA accumulator model. Hence, our analysis of the detection data focused on the Wald model[1], which we demonstrated in a parameter recovery study to be able to reliably recover the parameters explaining load effects, and to accurately estimate the uncertainty in these estimates (see supplementary materials).

We fit the Wald model to data from the conditions with high and low intensity stimuli simultaneously, assuming that the stimulus effect was potentially mediated by a rate difference and by a difference in threshold between blocks with different stimuli. We initially fit three models that differed only in their assumptions about the role of start-point noise, either that it was present and potentially different for the 3s and no-load conditions, that it was present but unaffected by load, or that it was absent. The latter model had 12 parameters: four threshold (B) parameters (none vs. 3s × LOW vs. HIGH accumulator), four rate (v) parameters (none vs. 3s × LOW vs. HIGH stimulus), and two each of non-decision (t_0) and response omission (p_f) parameters (none vs. 3s). The other two models added either one start-point variability parameter (A) or two (for none vs. 3s).

In contrast to the results for choice, the model with start-point variability was selected (DIC = -1553, w = .985), followed by the model where start-point variability differed with load (DIC = -1544, w = .015), with the model with no start-point noise clearly rejected (DIC = -1501, w = 0). Further analysis focuses on the lowest DIC model; as shown in supplementary materials

[1] Results for the initial LBA model fit are reported in supplementary materials, where it is shown that as for the choice data the LBA agreed with the Wald model in terms of its inferences about effects on parameters.

its fit was excellent. Although we confirmed a difference in start-point variability was not needed to model load, our parameter recovery study found that, in contrast to the other parameters, the start-point variability parameter produced relatively uncertain estimates. It was not needed to model choice, likely because in that case detection of the stimulus enabled participants to avoid sampling evidence before the stimulus appeared, which is thought to cause start-point variability (Laming, 1968). By definition, this is not possible in a detection task, so our results indicate this parameter should be included in model fits, and that this can be done so without compromising estimation of the other parameters.

Bayesian *p*-values indicated the same pattern as for the choice data. The response omission parameter was 1.1%, [.37, 1.79] higher in the 3s condition (6.1%, [5.6, 6.6] vs. 5.0%, [4.51, 5.50], $p = .002$), and non-decision time was 0.023s, [.013, .032] faster (0.152s, [.145, .158] vs. 0.175s [.168, .180], $p < .001$). The average response threshold was higher in the 3s condition (by 0.35, [.30, .40]: 1.17, [1.13, 1.20] vs. 0.82, [.78, .86], $p < .001$), the same was true individually for both high blocks (by 0.31, [.25, .36]: 1.13, [1.08, 1.18] vs. 0.83, [.78, .87], $p < .001$) and low blocks (by 0.39, [.33, .44]: 1.20, [1.15, 1.24] vs. 0.81, [.77, .86], $p < .001$). The average rate was clearly lower [.28, .46] in the 3s condition (3.10, [3.03, 3.19] vs. 2.72, [2.68, 2.80], $p < .001$), and again this was true for both high blocks [.43, .69] (3.32, [3.21, 3.44] vs. 2.80, [2.68, 2.85], $p < .001$) and low blocks [.06, .29] (2.91, [2.81, 2.99] vs. 2.72, [2.64, 2.81], $p < .001$). Thus, the increased response time under load was due to both higher thresholds and lower rates, but these effects were masked somewhat by non-decision time, which reduced the underlying 0.171s effect by around 13% to the observed 0.148s value.

We again checked these results using model-selection methods, by fitting six simplifications of the selected model. We also fit six corresponding simplifications of the model with no start-point noise, making a total of 14 models in the comparison. As for the choice data,

results of Bayesian *p*-values were confirmed, clearly selecting the model with start-point noise and allowing for load effects on rates, thresholds and non-decision time (see supplementary materials for details).

Choice vs. Detection. We extended our Bayesian *p*-values by comparing the size of the selected Wald models' load effects in detection and choice. There was little support for a decrease in the non-decision time load effect for detection compared to choice (M = .008s, p = .23, [-.03, .013]). However, there was some support for larger threshold effect in detection (M = .073, p = .05, [-.015, .16]) and strong support for a larger rate effect (M = .226, p = .001, [.096, .35]).

Discussion

A fundamental characteristic of human cognition is 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. ISO developed a novel metric, the DRT, for assessing cognitive workload in a variety of multitasking situations. DRT reaction time and miss rates are very sensitive to increases in cognitive workload; however, the precise reason is unclear – the differences could be due to changes in the rate of evidence accumulation and/or to a strategic adjustment in the amount of information required for a decision. This distinction could meaningfully change the approaches to applications of studying cognitive workload, such as alleviating driver distraction. 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. Large 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.

First, we found that in both choice and detection tasks the cognitive workload induced by the secondary task—counting backward by threes—reduced evidence accumulation rates. These results suggest information processing in choice and detection tasks 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 LBA (Brown & Heathcote, 2008) and the shifted Wald (Heathcote, 2004), both for choice and detection tasks. It confirms Strayer et al.'s (2011, 2015) interpretation of correlations between the DRT and effects of secondary tasks on driving performance as being mediated by limited-capacity attention. While this conclusion supports the current paradigm of cognitive workload as a competition for limited resources, this assumption has not been previously validated with the mathematical rigor of the methods described above.

Second, the choice task clearly reduced performance in the primary tracking task. There was a smaller reduction from detection of weak signals, with the ISO DRT having the smallest, but still reliable, impact. This is again consistent with all tasks drawing on a limited-capacity attention pool. The detection tasks were most sensitive to load effects regarding slowing, but less so for misses. They were also associated with larger load effects on the primary task, which were attenuated in the choice condition. Regarding model parameters, the small non-decision-time load effect was similar in detection and choice, but detection was more sensitive to the threshold increase and markedly more sensitive to the cognitive workload in accumulation rates. Overall, these results support detection rather than choice as being the best tool for measuring the cognitive workload effects of secondary tasks. They are also consistent with the ISO DRT with

an easily detected stimulus as being the best in practice regarding minimizing impact on the driving task. Therefore, simple and more easily automated secondary behavioral tasks can be used to detect differences in cognitive workload for goal directed behaviors with minimal impact upon primary tasks. However, the addition of the secondary task can have detectable effects upon the behaviors it attempts to measure.

Third, we accounted for response omissions by assuming a mixture of normal evidence accumulation and failures to encode the DRT stimulus. Ratcliff and Strayer (2014) took a different approach using a Wald model, assuming Gaussian trial-to-trial variability that sometimes results in negative rates and hence response omissions because 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 detection task, 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, and that it produces quite accurate and precise estimates of parameters relevant to cognitive workload effects with samples as small as 200 trials per participant, making it 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 a much wider range to behavioral tasks and psychological paradigms.

Fourth and 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 (~10-20%) slowing due to threshold increases and rate decreases. Another possibility is suggested by the results of Palada, Neal, Tay and Heathcote (In Press), which suggests that high cognitive loads could cause the stimulus encoding process to complete prematurely. In this view, response omissions can be seen as an extreme manifestation where

encoding failed completely, and so no evidence could be accumulated to initiate a response. The same mechanism may have been in part responsible for the reduced rate of evidence accumulation we observed under cognitive workload if weakened stimulus encoding causes a slower rate of accumulation.

Finally, we found an increase in both choice and detection thresholds due to cognitive workload, consistent with participants attempting to compensate for a known difficult goal by requiring more evidence for a response. Tillman et al. (2017) also found cognitive workload from 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 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. The divergence of results clearly indicates slowing in the DRT alone is not 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 in supplementary materials.

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 of what aspects of information processing are captured by the technique. In particular, the DRT is slowed due to a decrease in the rate of evidence accumulation. It is noteworthy that the DRT can be configured to be a portable system so that data can be collected anywhere. Given the ubiquity of multitasking in virtually all aspects of modern society, the DRT provides an exquisite tool for measuring capacity at its limits in a variety of everyday contexts. It is possible that future experimental manipulations will alter the

information processing dynamics underlying the DRT and the formal modeling described herein will assist in identifying if and when these changes occur. We foresee a large number of possibilities for this emerging methodology.

Context Paragraph

Strayer et al. (2015) utilized the International Standards Organization (ISO) standard Detection Response Task (DRT), amongst other converging behavioral, physiological, and neuroscience measures, to quantify the cognitive workload associated with using various mobile and in-vehicle technologies while driving. In some cases, interacting with these technologies produced slower and less accurate DRT responses compared to baseline driving and other activities that have been outlawed in many states, such as conversing on a cell phone (Strayer, Watson, & Drews, 2011). While these studies demonstrate that the DRT correlates with driving outcomes, the mechanisms through which the DRT captures the mental effort required in dual-tasking scenarios remains largely uninvestigated. Several competing proposed mechanisms would fundamentally change the current approach to the study of cognitive workload, and subsequent attempts to alleviate it in applied settings. In a collaborative effort to apply rigorous quantitative methods to the cognitive processes underlying driver workload, we developed a research program together with the creators of the Linear Ballistic Accumulator model (Brown & Heathcote, 2008) to address the interaction of low-level perceptual and motor processes with higher order cognitive processes in producing variable driving performance. Also, by providing an example of our model's application, the dataset, and our custom software (https://osf.io/e8kag/) based in the open-source language and statistical computing environment R (R Core Team, 2016), we have provided a template for future modeling of cognitive workload measurements in other naturalistic settings.

References

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.

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.

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.

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., 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.

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.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *reason*, *4*(2), 61-64.

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. (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., & 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.

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.

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

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.