Dynamic Workload Measurement and Modeling: Driving and Conversing

Spencer C. Castro[1], Andrew Heathcote[2], Joel M. Cooper[3], and David L. Strayer[4]

[1] Management of Complex Systems Department, University of California, Merced, scastro39@ucmerced.edu
[2]School of Psychology, University of Newcastle, Australia, andrew.heathcote@newcastle.edu.au
[3] Red Scientific Inc., joel@redscientific.com
[4]Department of Psychology, University of Utah, david.strayer@utah.edu

Wordcount (excluding abstract and references) = 5585

This research was supported in part by the National Science Foundation Graduate Research Fellowship Program, the AAA Foundation for Traffic Safety, and the Australian Research Council (Grant DP160101891). Correspondence concerning this article should be addressed to Spencer C. Castro, Management of Complex Systems Department, University of California Merced, Merced, CA 95343. E-mail: scastro39@ucmerced.edu. Data and analyses can be found on the Open Science Framework at osf.io/4a9xb.

Abstract

Tillman et al. (2017) used evidence-accumulation modeling to ascertain the effects of a conversation (either with a passenger or on a hands-free cell phone) on a drivers' mental workload. They found that a concurrent conversation increased the response threshold but did not alter the rate of evidence accumulation. However, this earlier research collapsed across speaking and listening components of a natural conversation, potentially masking any dynamic fluctuations associated with this dual-task combination. In the present study, a unique implementation of the Detection Response Task was used to simultaneously measure the demands on the driver and the non-driver when they were speaking or when they were listening. We found that the natural ebb and flow of a conversation altered both the rate of evidence accumulation and the response threshold for drivers and non-drivers alike. The dynamic fluctuations in cognitive workload observed with this novel method illustrate how quickly the parameters of cognition are altered by real-time task demands.

Public Significance Statement: This study presents a novel method for measuring *and* modeling the dynamic fluctuation in workload of in-person and cell-phone conversations of both the driver and the non-driver. Both interlocutors exerted more mental effort while speaking than listening and the effects of the conversation were additive with the driving task. Our modeling suggests that the increased workload associated with conversing whilst driving is due to a decrease in rate of evidence accumulation and an increase in response caution.

The real-world tasks of driving an automobile and engaging in a conversation with an interlocutor are continuous, and each task results in dynamic fluctuations in the effort required to maintain acceptable levels of performance. Driving performance fluctuates with the difficulty of the driving task (Teh, Jamson, Carsten, & Jamson, 2014). The demands of a conversation fluctuate between speech comprehension, which is easier, and speech production, which is harder, (e.g., Lee, Cerisano, Humphreys, & Watter, 2017, but see Kubose et al., 2006). When driving and conversing are performed concurrently, they compete for limited capacity attention (Kahneman, 1973). For example, driving performance degrades as the difficulty and complexity of the conversation increases (McKnight, & McKnight, 1993; Nunes & Recarte, 2002). Similarly, a conversation degrades as the demands of driving increase (Nunes & Recarte, 2002; Drews, Pasupathi, & Strayer, 2008).

Tillman, Strayer, Eidels, and Heathcote (2017) measured the cognitive workload of a dyad engaged in a natural conversation. They contrasted an in-person conversation (i.e., between a driver and a passenger in a vehicle) with a hands-free cell phone conversation in which the driver and non-driver were in different physical locations. To obtain measures of workload, they implemented a detection response task (DRT; International Standards Organization, ISO 17488, 2016), which has been shown to be sensitive to cognitive workload (e.g., Castro, Cooper, & Strayer, 2016; Cooper, Castro, & Strayer, 2016).every 3-5 seconds, "yoked" DRT devices (one fitted to the driver and one fitted to the non-driver) flashed a light in the peripheral field of view of the left eye of each member of the conversational dyad. Both the driver and non-driver responded separately to the onset of the light by pressing a microswitch attached to their finger.

Tillman et al., (2017) found that DRT responses were faster when the *driver* was not conversing (i.e., driving only) than when they were also conversing in person or over a cell phone.

The response time (RT) for the latter conditions did not differ, in line with previous research (e.g., Nunes & Recarte, 2002; Strayer & Johnston, 2001; Strayer, Drews, & Johnston, 2003). The *non-driver's* responses were faster than the driver and, like with the driver, were equivalent for passenger and cell phone conversations. The fact that RT was elevated for the driver suggests that the driving task and the conversation task compete for limited attentional resources. Tillman et al., (2017) modeled the driver's DRT performance using an evidence-accumulation model that enables measurement of the mean rate of evidence accumulation (i.e., drift rate), the threshold amount of evidence required to trigger a response (i.e., response caution), and non-decision time (i.e., the time to complete perceptual encoding and motor response processes) (Brown & Heathcote, 2008; Ratcliff & McKoon, 2008). A single-bound diffusion model (Heathcote, 2004; Logan, Van Zandt, Verbruggen, & Wagenmakers, 2014) was fitted to the driver's DRT responses and found that the workload effect was due to an increase in a participant's response thresholds. There was scant evidence that the rate of evidence accumulation for the driver changed with the addition of the conversation task.

This pattern is surprising because the driving and conversation tasks seem to compete for limited attentional resources (e.g., Kahneman, 1973), a pattern that should theoretically impact the rate of evidence accumulation. Under this logic, the rate of evidence accumulation should decrease when a participant divides their attention between two attention-demanding tasks.

As evidence-accumulation modeling originates from the decision-making literature (e.g., Ratcliff & Rouder, 1998), the theoretical framework of this approach suggests two types of mental processes. Bargh and Chartrand (1999) listed the examples of these processes as "conscious—nonconscious, controlled—automatic, explicit—implicit, systematic—heuristic", where generally one refers to willfully regulating behavior and the other to a process that the individual is less

aware of. In this framework, Tillman et al., (2017) suggest that only the certainty, or cautiousness, with which we make a response matters in terms of cognitive workload.

Several of the same decision-making frameworks have been applied to describe the mechanisms resulting in multitasking limitations (e.g., Kahneman, 1973; Navon & Gopher, 1979). However, multitasking requires allocating attention among two or more goals (e.g., Braver, 2012; Wickens & McCarley, 2019). Researchers argue that this process involves a fundamentally different mechanism than maintaining the effort to accomplish one goal (e.g., Howard, Evans, Innes, Brown, & Eidels, 2020; Norman & Shallice, 1986). Using this framework, multitasking can be evaluated using a dual-process model of performance with the mean rate of evidence accumulation (i.e., drift rate) reflecting one process and the threshold amount of evidence required to trigger a response (i.e., response caution or bias) reflecting the other process.

However, an important limitation of Tillman et al., (2017) is that they averaged over the speech comprehension and speech production components of the conversation, possibly weakening any effects of workload on the rate of evidence accumulation. As noted above, speech comprehension and production place different demands on an interlocutor, particularly if they are concurrently driving an automobile (e.g., Strayer et al., 2015). In the conversational dyad, there should be a reciprocal pattern of workload such that workload is higher for the driver when they are speaking than when they are listening. By contrast, the workload of the non-driver should be higher when the driver is listening (and the non-driver is talking) and lower when the driver is talking (and the non-driver is listening). Therefore, these effects of speech should not be collapsed across due to their differential impact on workload. The current study uses the dual-DRT configuration developed by Tillman et al., (2017) to measure and model the performance of *both* the speaking and listening of the driver and non-driver as they engage in a naturalistic conversation

in-person or remotely over a hands-free cell phone. To discriminate between fluctuations in speaking and listening, microphones were attached to each of the DRT devices (i.e., one for the driver and one for the non-driver) and the audio was used to trigger a code for who was speaking and who was listening.

Behaviorally, we predict that DRT RT will be longer, and the probability of responding (i.e., hit rate) lower, for the driver than for the non-driver, reflecting the added load associated with driving. We also predict that RT will be higher and hit rate lower when the participant is speaking than when they are listening. Moreover, we predict that the pattern of DRT data for the driver and non-driver will mirror one another.

Unlike Tillman et al., (2017), who modeled only the driver's DRT performance collapsed over speaking and listening using the single-bound diffusion model, we separately modeled cell phone and passenger conversations for the passenger. Tillman et al. focused only on response times and did not take into account failures to respond to the DRT stimulus (omissions) as they occurred at a low rate when in the conversation conditions (~4%) and not at all in their condition with no conversation. In the experiment reported here, omissions occurred at a higher rate and differed more markedly across conditions, suggesting that they could not be ignored. Hence, we used Damaso et al.'s, (2021) Linear Ballistic Accumulator with Omissions (LBAO) model to provide a simultaneous account of both RT and omissions.

Like the diffusion model, the LBAO has parameters for response caution (B), the mean rate of evidence accumulation (v), and perceptual encoding and motor response production time (t_0). We predict that the rate of evidence accumulation and the threshold for responding to the DRT will be modulated by the dynamic fluctuations in workload associated with the conversation, so conversational turn-taking will result in shifts in both threshold and rate parameters.

Additionally, the LBAO has a parameter for trial-to-trial variability in drift rate (*sv*), which follows a normal distribution. When rate variability causes a sufficiently small rate to be sampled an omission can occur because the threshold amount of evidence is not accumulated before the maximum time allowed for a DRT response (3 s). Ratcliff and Strayer (2014) used the same mechanism in the diffusion model to account for omissions, but this model could not distinguish between rates and threshold and had no closed-form likelihood, making it difficult to fit. The LBAO overcomes both of these limitations, enabling an understanding of the effects of dynamic fluctuations of cognitive workload on omissions while also disentangling rate and threshold effects. We predict that manipulations that reduce capacity will also increase omissions by making small rates more likely.

Method

Transparency and Openness

Data

The data for the discussed studies are available with the link posted here and in the Author Note: https://osf.io/4a9xb/

Analytic methods

The analytic code needed to reproduce analyses is available and the link to access this information is provided in the Author Note with behavioral analyses within *Final_Analysis cop.zip* and modeling analyses within *LBAO_Modeling_ConvoDrive copy.zip*.

Materials

Certain materials, such as the DS-600 DriveSafety™ driving simulator, the driving scenario, and the Detection Response Task devices (ISO 17488, 2016) are not publically available. However, DRT devices and standards can be purchased from https://redscientific.com/ and the

International Organization for Standardization 17488. Methods for the LBAO can be found in Damaso et al. (2021) and its supplementary materials.

Participants

Forty-four participants (23 Female, age: M = 21.1, SD = 3.4) were recruited in 22 dyads from University of Utah undergraduate psychology courses. Information on race, ethnicity, or socioeconomic status was not collected. Participants received course credit as compensation for completion of the one-hour study. Dyads were required to know each other in order to facilitate naturalistic conversation.

Materials

A DS-600 DriveSafety™ driving simulator provided the experience of driving an automatic shifting compact passenger sedan. DriveSafety software was utilized to program a 19-mile driving scenario, which included two- and three-lane divided highways with speed limits between 55 and 65 miles per hour and moderate traffic. The other simulated vehicles changed speed and lanes to create *irregular-flow* traffic (Drews, Pasupathi, & Strayer, 2008), which simulated realistic traffic. Participants drove for approximately 15-minutes in each block. The driving environment, the familiar dyads, and the naturalistic conversation topics were all designed to simulate natural conversations while driving on a highway.

Separate vibrotactile Detection Response Task devices (ISO 17488, 2016), one for the driver and the other for the non-driver, were used throughout the experiment. Following ISO 17488 (2016) guidelines, a small vibrating motor was attached to each of the participants' left collarbone at the base of the neck. The DRT onset occurred pseudo-randomly every 3-5 seconds, and participants responded with a small button attached to their right thumb. The vibration stopped when the participant pressed the button or after one second had elapsed. A dedicated

microprocessor recorded millisecond-accurate responses. Microphones were utilized to determine whether a DRT stimulus occurred while the driver or non-driver was speaking, or during silence.

Table 1. *Experimental Design for a 2 between (Condition) X 2 within (Role) X 3 within (Speaker) Repeated Measures Linear Mixed-Effects Model.*

Condition	**Role**	**Speaker**
Randomly Assigned (between dyads)	Within Participant (2x15-minute blocks)	Within Participant (unbalanced time)
Cell Phone	Driver	None Driver Other
	Other	None Driver Other
In Person	Driver	None Driver Other
	Other	None Driver Other

Procedure

Dyads were randomly assigned to either the cell phone or passenger conversation condition and participants were randomly assigned to the driver or the other role first. Both participants in the dyad completed the driving and non-driving roles, but they only participated in the cell phone or passenger conversation condition (i.e., in-person vs. cell phone conditions was a between-subjects factor, see Table 1). Participants first performed a 5-minute practice block where one participant sat in the driver's seat and the other participant either sat in the passenger seat or at a remote location. Both participants responded to the DRT for the 5 minutes and the participant in the driver's seat drove the vehicle. Participants then switched roles and the procedure was repeated. After the practice drives, participants selected 10 topics from a list of 20 *conversation starters* listed by Psychology Today (Barreca, 2017; Prompts are listed in Appendix A).

Participants held a conversation for 15 minutes while driving and either sitting in the passenger seat or at the remote location using a hands-free cell phone. Participants then switched places and completed another 15 minutes of conversation.

Measures. DRT RT and Hit Rate (HR) was recorded for both participants. The DRT stimulus presentation to the driver and the non-driver were independent. Following ISO guidelines (ISO 17488, 2016), anticipatory responses shorter than 100 milliseconds (0.09%) were excluded from statistical analysis. Driving performance measures included speed variability, Root Mean Squared Error (RMSE) from the speed limit, and lateral lane deviation.

Results

Differences in conditions between having a conversation with a passenger and having a conversation with a remote participant via a hands-free cell phone were tested in R (R Development Core Team, 2018). The lme4 package (Bates, Maechler, Bolker, & Walker, 2015) was used to create linear mixed-effects models (LMMs) with fixed effects of condition (2: cell phone vs. in-person), the participant role (2: driver vs. non-driver), and speaking demands (3: no talking, listening, and speaking) fully crossed. We report Type II Wald chi-square tests of differences in RT, HR, and RMSE across conditions; 95% confidence intervals are reported in square brackets. In all cases likelihood ratio tests selected random slopes for the effect of drive on participants and random intercepts for participants to account for the experimental design of two drives per participant. Additionally, we included block as a covariate to account for any effects of learning or fatigue.

Behavioral Measures

RT

Statistical analyses were performed on log-transformed RTs but are not transformed in Figure 1 for clarity. Conversing increased RT over no talking by 39 ms [23, 54], $\chi^2(2) = 850.67$, $p < .001$ and drivers responded more slowly than non-drivers by 89 ms [75, 102], $\chi^2(2) = 8.13$, $p = .004$. RT for participants in the cell phone condition did not differ significantly from when participants conversed in person, $\chi^2(1) = 2.76$, $p = 0.12$.

Speaker interacted with participant role, $\chi^2(2) = 356.96$, $p < .001$, and with condition $\chi^2(2) = 37.81$, $p < .001$. In the in-person condition, the increase in RT from no talking to the driver speaking was smaller for the *non-driver* (27 ms, 95% CI [23, 31]) than for the *driver* (102 ms, [98, 106]). By contrast, the increase in RT between no talking and the passenger speaking for the *non-driver* (103 ms, [99, 107]) was greater than for the *driver* (14 ms, [4, 23]). In the cell phone condition, the increase in RT from no talking to the driver speaking was smaller for the *non-driver* (27 ms, [23, 31]) than for the *driver* (48 ms, [38, 60]). By contrast, the increase in RT between no talking and the passenger speaking for the *non-driver* (102 ms, [99, 107]) was greater than for the *driver* (71 ms, [4, 23]).

A three-way interaction, $\chi^2(2) = 9.33$, $p = .009$, was driven by the differences in RT between the cell phone and in-person conditions for the *driver* when they were listening to the non-driver speak. This suggests that the driver found it more difficult to listen to the other talking in a cell phone conversation than to an in-person conversation (see Figure 1).

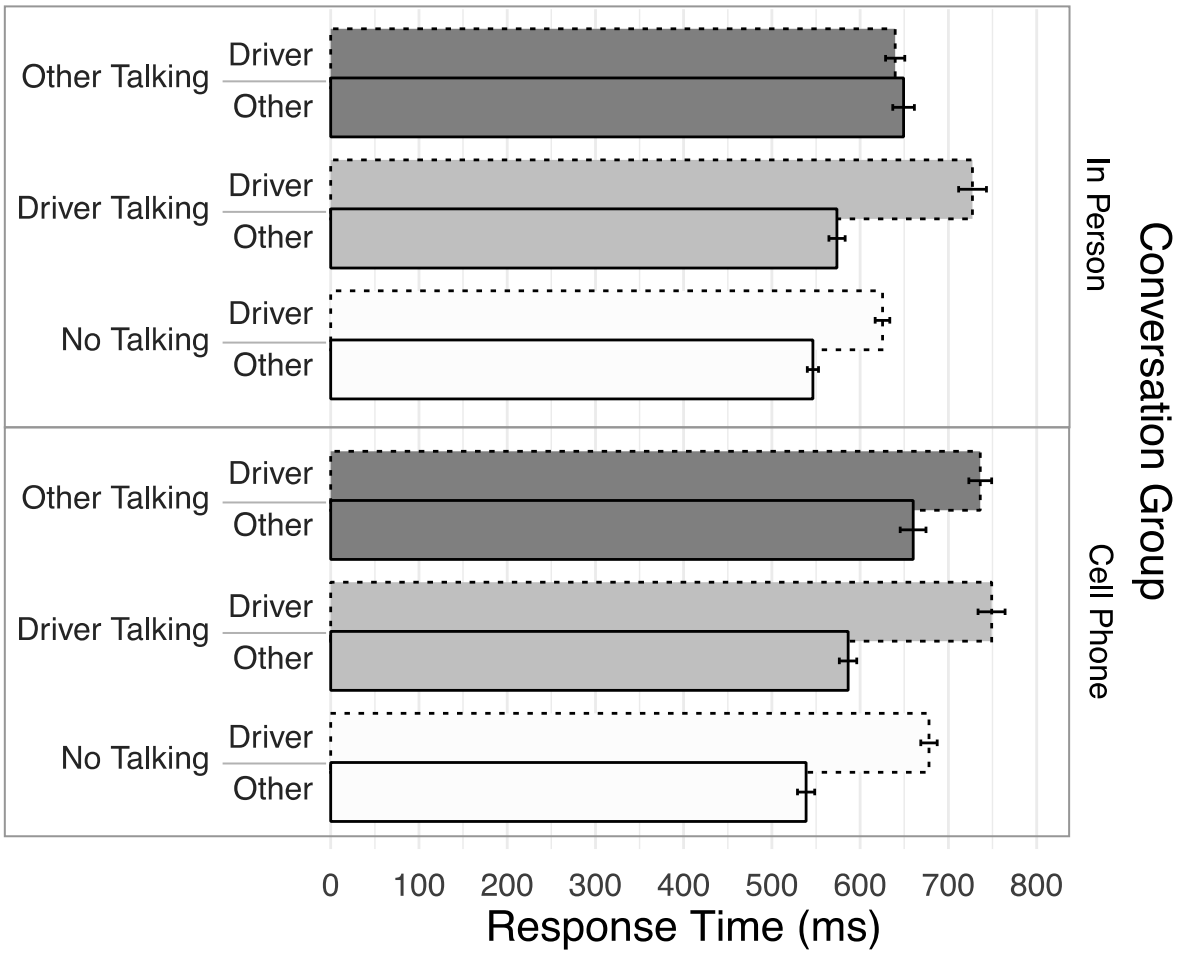

Figure 1. Response times of the driver and other participant during a 15-minute drive, according to who was talking and who was listening. 95% Confidence Intervals were calculated utilizing the Cousineau-Morey method for repeated-measures designs (Cousineau, 2005; Morey, 2008; Baguley, 2012).

Hit Rate

A binomial LMM with a probit link was fit by maximum likelihood using Laplace Approximation (see Figure 2 for means and 95% CIs). Conversing decreased HR over no talking by 4.21% [3.21, 6.54], $\chi^2(2) = 850.67$, $p < .001$, and drivers responded less often than non-drivers by 1.32% [.75, 1.52], $\chi^2(2) = 8.13$, $p = .004$. The effect of condition (i.e., cell phone vs. in-person) on HR was significant, $\chi^2(1) = 10.39$, $p = .001$.

The effect of condition interacted with who was responding to the DRT, $\chi^2(1) = 153.04$, *p*

< .001. Who was speaking also interacted with who responded to the DRT, $\chi^2(1) = 39.77$, $p < .001$. The speaker was part of a significant interaction with the responder, $\chi^2(1) = 72.22$, $p < .001$ (see Figure 2). As with the RT data, this suggests that the driver found it more difficult to listen to a cell phone conversation than to an in-person conversation.

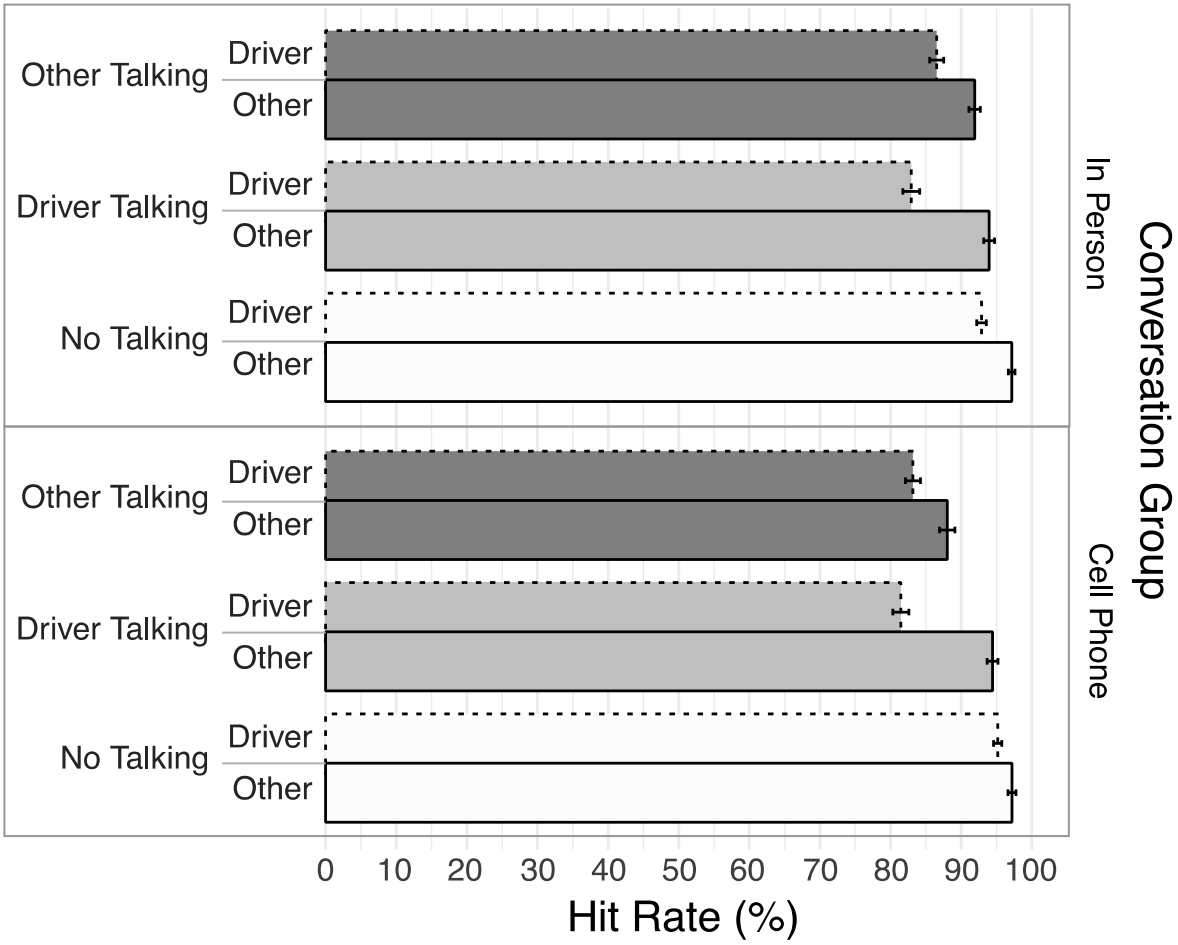

Figure 2. Hit Rate of the driver and other participant during a 15-minute drive, according to who was talking and who was listening. 95% Confidence Intervals were calculated utilizing the Cousineau-Morey method for repeated-measures designs (Cousineau, 2005; Morey, 2008; Baguley, 2012).

Driving Performance

Drivers exhibited greater lateral steering deviation when talking (M = .37 m, [.16, .58]) than when listening (M = .30 m, [.17, .43], $\chi^2(1) = 73.23$, $p < .001$). Additionally, drivers in the

cell phone condition produced slightly smaller steering deviation (M = .31 m, [.17, .45]) than drivers in the passenger condition (M = .34 m, [.16, .52]), $\chi^2(1) = 7.32$, $p = .04$. A similar main effect was found for compliance with the speed limit (RMSE) between no talking, listening, and speaking conditions $\chi^2(1) = 11.45$, $p = .021$. However, the difference between speaking and listening conditions did not achieve significance, $\chi^2(1) = 1.45$, $p = .09$. Finally, participants drove more slowly overall when talking on a cell phone (M = 62.06 mph, [59.80, 64.32]) than when conversing with a passenger (M = 64.2 mph, [62.30, 66.10]) $\chi^2(1) = 9.45$, $p = .03$. Table 2 reports individual condition means and standard deviations.

Table 2. *Means and Standard Deviations of Driving Performance for the Cell Phone and In Person conditions when nobody is speaking (i.e., None), when the driver is speaking (i.e., Driver), and when the other participant is speaking (i.e., Other).*

	In Person			**Cell Phone**		
	Speaker			**Speaker**		
	None	Driver	Other	None	Driver	Other
	M (CI)	*M* (CI)	*M* (CI)	*M* (CI)	*M* (CI)	*M* (CI)
Steering Deviation (RMSE)	0.30 (.13)	0.37 (.21)	.31 (.18)	.28 (.12)	.34 (.18)	.31 (.14)
Speed Variability (RMSE)	2.5 (1.2)	3.5 (1.4)	2.9 (1.1)	2.1 (.81)	3.1 (1.0)	2.4 (1.2)
Average Speed (mph)	66.3 (5)	63.2 (6)	64.3 (5)	64.3 (4)	60.5 (5)	61.4 (5)

Note. RMSE = root mean squared error; mph = miles per hour.

Modeling Approach

Damaso et al., (2021) defined two types of omissions that occur because of drift rate variability. “Intrinsic” omissions occur because the sampled rate for a trial can be negative, and so threshold will never be reached. “Design” omissions occur when a positive drift rate is sampled that is too small to reach the threshold in the 3 s response window. Drift rates in the LBA vary according to a normal distribution, and so the probability of an intrinsic omission in

speaking-demand condition *i* is $p_I = \Phi(v_i/sv_i)$, where $\Phi(x)$ is the integral of a normal distribution with mean *x* and a unit standard deviation from $-\infty$ to zero, and v_i and sv_i are the rate distribution's mean and standard deviation. Design omissions speaking-demand condition *i* occur with probability $p_{Di} = 1 - F(t = 3 \mid A_i, b_i, v_i, sv_i)$, where *F*() is the cumulative distribution function for an LBA accumulator (see Brown & Heathcote, 2008), A_i is the range of trial-to-trial variability in the starting point of accumulation, and b_i the threshold in condition *i*. Combining the two types the total omission probability is $p_{Ti} = p_{Ii} + (1 - p_{Ii})\, p_{Di}$, where the subscript T indicates that the combined value is a function of task-related factors as it depends on the same parameters that determine DRT responses. Note that no extra parameters need to be estimated to produce predictions about task-related omissions.

Damaso et al., (2021) also included an extra parameter, p_C, to allow for "contaminant" omissions with a different origin to task-related omissions. This idea was drawn from Castro et al. (2019), who attributed such omissions to a failure to encode the stimulus. They estimated different values of p_C to account for increased DRT omissions when participants performed a secondary task (counting backward). By itself this approach is essentially only descriptive as there is no relationship between response and omission processes. We preferred Damaso et al.'s approach as it links the two and so is naturally able to account for correlations between these two performance measures. Damaso et al. included both task-related and contaminant omissions because they were requried to account for a few participants with high overall omission rates but treated their probability as an individual difference variable that does not change with secondary-task workload.

We made the same assumption about contaminant probability with respect to speaking-demand conditions, so only one extra parameter is estimated, and the overall probability of

omissions in condition *i* is $p_{Oi} = p_C + (1 - p_C)p_{Pi}$. Hence, any differences in omission rates between conditions are entirely accounted for by the parameters of the evidence accumulation process, and so we focus on these parameters in our analysis. Like Damaso et al. (2021), we found that contaminant omissions varied widely over individuals, being less than 5% for most but up to almost 15% for a few participants.

Modeling Results

The DMC software (Heathcote et al., 2018) was used to fit models in a Bayesian manner separately to each combination of role (driver vs. passenger) and speaker-type condition (in person vs. cell phone). In each case we fit 8 models that allowed various combinations of thresholds (*B*) rates (*v*), rate standard deviations (*sv*), and non-decision time (t_0), were allowed to differ with the three speaking-demand conditions (none, driver, other). To make models that allowed threshold to vary with speaking-demand condition identifiable the threshold in the no-talking condition was fixed to one. Similarly, for models that assumed the same threshold for all speaking-demand conditions its value was set at one.

In every case, the best model according to the DIC model selection criterion included an effect of speaking-demand condition on the threshold parameter and both drift rate parameters but dropped an effect on non-decision time (see Table 3). This model was used in all further analyses.

Table 3. *The difference between Deviance Information Criterion (smaller values indicate a better tradeoff between goodness-of-fit and model complexity) relative to the best model (i.e., the DIC value for the best model in each row is subtracted from the values for all models in the row, so the best model has an entry of zero). The models either allowed thresholds (B), rates (v), rate standard deviations (sv) or non-decision time (t0) to vary over the three speaking-demand conditions. Models were fit to each condition and pair member role separately.*

Driver with Passenger	*Bvsv*	*Bvt*$_0$*sv*	*vt*$_0$*sv*	*Bv*	*Bvt*$_0$	*vt*$_0$	*vsv*	*v*
	0	126	204	296	418	600	668	1326
Passenger with Driver	*Bvsv*	*Bvt*$_0$*sv*	*Bv*	*vt*$_0$*sv*	*Bvt*$_0$	*vt*$_0$	*vsv*	*v*
	0	84	298	331	410	557	669	1329
Driver with Cell Phone	*Bvsv*	*Bvt*$_0$*sv*	*vt*$_0$*sv*	*vsv*	*Bvt*$_0$	*Bv*	*vt*$_0$	*v*
	0	4	111	509	855	1016	1225	1729
Non-driver with Cell Phone	*Bvsv*	*Bvt*$_0$*sv*	*Bv*	*vt*$_0$*sv*	*Bvt*$_0$	*vt*$_0$	*vsv*	*v*
	0	134	237	272	428	573	676	1313

Note. For a discussion on the Deviance Information Criterion, see Spiegelhalter, Best, Carlin, & Van Der Linde, (2002).

Parameter Tests

We report parameter estimates from the best model as posterior medians with 95% credible intervals (in square brackets) and *p* value indicating the fixed-effect probability that one parameter is greater than another with small *p* values supporting a difference.

Driver Responses with a Passenger. The response threshold (*B*) remained fixed at 1 while there was silence in order to make the model identifiable. We found that the response threshold (*B*) was larger for talking than listening (.74 [.68, .80] vs .58 [.52, .63], respectively, $p < .001$). The mean rate (*v*) decreased from silence to listening (3.52 [3.44, 3.61] vs 2.41 [2.24, 2.60], respectively, $p < .001$) and from silence to talking (3.5 [3.44, 3.61] vs 2.49 [2.34, 2.66], respectively, $p < .001$). There was little evidence for a difference in mean rate (*v*) between listening and talking (2.41 [2.40,2.60] vs 2.49 [2.33, 2.66], respectively, $p = .21$).

Passenger Responses with a Driver. With the silent responses fixed at 1, threshold (*B*) increased from listening to talking (.58 [.52, .63] vs .73 [.68, .80], respectively, $p < .001$). The mean rate (*v*) decreased from silence to listening (3.52 [3.44, 3.61] vs 2.41 [2.24, 2.60], respectively, $p < .001$) and from silence to talking (3.5 [3.44, 3.61] vs 2.49 [2.34, 2.66],

respectively, $p < .001$). There was little evidence for a difference in mean rate (v) between listening and talking (2.41 [2.24,2.60] vs 2.49 [2.34, 2.65], respectively, $p = .22$).

Driver Responses with a Cell Phone. Again, the response threshold (B) remained fixed at 1 while there was silence in order to make the model identifiable. There was no evidence to suggest that listening had a higher threshold than talking (1.85 [1.60, 2.12] vs 1.82 [1.55, 2.14], respectively, $p = .44$). The mean rate (v) increased from silence to listening (4.83 [4.66, 5.01] vs 6.33 [5.65, 7.03], respectively, $p < .001$) and talking (4.83 [4.66, 5.01] vs 5.93 [5.28, 6.63], respectively, $p < .001$). There was little evidence for a difference between listening and talking (6.33 [5.65, 7.03] vs 5.93 [5.28, 6.63], respectively, $p = .19$)

Non-Driver with a Cell Phone. With the silent responses fixed at 1, threshold (B) increased from listening to talking (.65 [.55, .74] vs .83 [.72, .97], respectively, $p < .01$). The mean rate (v) decreased from silence to listening (3.49 [3.40, 3.59] vs 2.40 [2.20, 2.61], respectively, $p < .001$) and talking (3.49 [3.40, 3.59] vs 2.50 [2.26, 2.77], respectively, $p < .001$). There was little evidence for a difference between listening and talking ([2.20, 2.61] vs 2.50, [2.64, 2.77], respectively, $p = .21$).

Contributions to Workload

Workload is considered a multi-dimensional concept that one representative measure will fail to capture (Gopher & Donchin, 1986). However, several mechanisms of performance and effort requirements can be discussed within the context of goal-directed behavior. For example, the process of changes in workload measurements can be consciously mediated, with participants slowing in their responses or failing to respond because they deliberately respond more cautiously with higher workloads in the primary task. If the driver maintains separate resource pools for the primary and secondary tasks, like in some resource theories of attention (e.g., Wickens, 2008),

then a strategic increase in response certainty (i.e., caution) would be responsible for slower DRT responses. If the tasks required separate resources, dual-task costs would not be observed at all. In an applied setting, a driver could prioritize reacting to traffic changes over DRT responses when they perceive increased driving difficulty. Previous research demonstrates a strong correlation between the DRT and self-report measures of subjective workload (Strayer et al. 2013), such as the NASA Task Load Index (Hart and Staveland 1988).

However, reductions in a unitary pool of resource may occur due to both tasks in a more traditional theory of attention (Kahneman, 1973), especially when there is an implicit task priority. In this case, the primary task would receive the resource necessary to complete the task to the operator's perceived ability and whatever is left would be allocated to the second task. If the resources allocated did not match the required workload, processing of that task would slow. In our modeling terms, this would be a slowing of the rate of evidence accumulation toward a response.

To our knowledge, this is the first study to measure *and* model the natural and dynamic ebb and flow of mental workload of both interlocutors in a conversational dyad, as most studies block the experimental conditions and only obtain measurements from one participant. We observed a reciprocal tradeoff in workload for the conversational dyad such that when one participant was speaking their workload, as inferred from the DRT, was higher than when they were listening. An inverse pattern was observed for the other participant (e.g., higher workload when the other participant was listening than when the other participant was talking). Moreover, the driving task showed an additive relationship with the conversation (i.e., higher for the driver than the non-driver), suggesting that the driving and the conversation tasks compete for the same limited capacity resources. The data help to explain why a conversation can lead to driver-

restricted attention (e.g., Regan, Hallett and Cordon, 2011). This impairment is most apparent with cell phone conversations due to the compensatory factors associated with passenger conversations (e.g., Drews, Pasupathi, and Strayer, 2008).

Our model selection indicates that differences in threshold and rate (both mean and variability) play a significant role in explaining the variations in workload. To quantify their relative importance, we systematically held constant the effect of each parameter by setting the parameter to its average value across conditions while leaving the other parameters at their estimated values and simulated DRT data to determine the reduction in the model's ability to account for the workload differences (Strickland, Loft, Remington & Heathcote, 2018). The reduction in variance accounted for was computed using the following equation:

$$Percent\ of\ Variance\ = (1 - \left(\frac{Bvsv\ Model\ -\ Fixed\ Parameter\ Model}{Bvsv\ Model}\right) * 100)$$

Driver Responses with a Passenger. We found that for drivers with a passenger, RT slowed 79 [43, 115] milliseconds when talking compared to listening, but that this effect disappeared in the simulated data when the threshold (*B*) was fixed at its average (5 [-169, 179] ms). When mean rate (*v*) was fixed, the effect decreased by 21 milliseconds to 59 [-43, 161] milliseconds. With fixed *B*, the model lost 93.67% of the effect while with fixed mean *v*, the model lost 26.58% of the effect.

Passenger Responses with a Driver. We found that the effect of talking compared to listening (79 [43, 115] ms) disappeared for the simulated data when the threshold (*B*) was fixed at its average (-58 [-153, 37] ms)). When mean rate (*v*) was held fixed, the effect was hardly reduced (75 [12, 137] ms). The effect of removing *B* accounted for an increase in the effect of 73% while removing mean rate (*v*) decreased 5% of the effect.

Driver Responses with a Cell Phone. We found that for drivers talking on a cell phone, RT slowed 74 [33, 115] milliseconds when talking compared to silence, but that this effect disappeared for the simulated data when the threshold (*B*) was fixed at its average (5 [-169, 179] ms). When mean rate (*v*) was fixed, the effect decreased by 21 milliseconds to 59 [-43, 161] milliseconds. With fixed *B*, the model lost 85.64% of the effect while with fixed mean *v*, the model lost 26.58% of the effect.

Non-Driver with a Cell Phone. We found that for drivers talking on a cell phone, RT slowed 93 [45, 141] milliseconds when talking compared to silence, but that this effect disappeared for the simulated data when the threshold (*B*) was fixed at its average (15 [-121, 149] ms). When mean rate (*v*) was fixed, the effect decreased by 11 milliseconds to 82 [-13, 181] milliseconds. With fixed *B*, the model lost 83.87% of the effect while with fixed mean *v*, the model lost 11.83% of the effect.

Discussion

We found that the workload of the driver and non-driver traded off in a naturalistic conversation. Overall, the passenger's workload was lower than that of the driver. The fact that RT was elevated for the driver suggests that the driving task and the conversation task competed for limited attentional resources. The reciprocal pattern observed when the dyad was conversing (e.g., higher for the driver and lower for the non-driver when the driver was speaking than when the driver was listening) demonstrates the complexity of measuring the cognitive workload of conversations while driving.

The DRT data were modeled using Linear Ballistic Accumulation with occasional response Omissions (LBAO). According to the *Deviance Information Criterion*, the model omitting t_0 (i.e., *Bvsv*) best fit the data, showing that the perceptual encoding and motor-response

parameters were not necessary to differentiate between the experimental conditions. The primary factor differentiating a driver talking compared to not talking was the threshold to a response (the model with fixed *v* accounting for 93% of the effect when with a passenger and 85% when talking on a cell phone), with the evidence accumulation rate (i.e., the model with fixed *B*) accounting for about 26% of the effect.

Importantly, the LBAO modeling shows that the division of attention between driving and conversing reduces the rate of evidence accumulation. The pattern varies slightly for in-person and cell phone conversations, but in both cases the workload differences are the result of changes in *both* the response threshold and the rate of evidence accumulation. This pattern differs from that of Tillman et al., (2017), in which dual-task costs were attributed solely to changes in response threshold. The current data suggests that aggregating over listening and speaking may have masked any effects of workload on the rate of evidence accumulation.

It is worth considering three possible outcomes of the LBAO modeling and how they might inform our theoretical understanding of dual-task performance. One possibility is that the addition of a secondary task increases the response threshold, but leaves the rate of evidence unchanged (e.g., Tillman et al., 2017). Essentially, dual-task performance is slowed under this scenario because of response caution. Another possibility is that the addition of a secondary task decreases the rate of evidence accumulation but leaves the response threshold unchanged. Essentially, dual-task performance is slowed under this scenario because the bandwidth of information processing for each task has been reduced by a finite resource allocation policy (e.g., Kahneman, 1973). Finally, the addition of a secondary task could *increase* the response threshold and *decrease* the rate of evidence accumulation. This latter possibility is what was observed in the current study. Dual-task performance was altered because of response caution and a splitting of the information

processing bandwidth.

Previous research can account for the presence of these parameters. For example, Drews, Pasupathi, & Strayer, (2008), demonstrated that passengers could modify their behavior when perceiving an especially high workload driving environment. This could be a source of differences in model parameters between the non-drivers, where a majority of the conversational differences are captured by prioritization decisions (i.e., the threshold parameter), and the drivers, where a larger proportion of the effect is accounted for by limited capacity (i.e., the rate parameter).

Our research was not sufficiently powered to examine differences in the nature of the conversation (e.g., neutral vs emotional), the degree of familiarity of the dyads, or individual differences in capacity. We posit that future research using the dual-DRT methodology will help to shed light on how workload is modulated by these factors, as literature suggests that they have an impact on aggregate performance (e.g., Hickman, Soccolich, Fitch, & Hanowski, 2015)

Other limitations include an inability with the current study's design to meaningfully address questions of individual differences within dyad interactions. Although the study recruitment approach helped to ensure that all the participants previously knew the other member of the dyad, the degree of familiarity was not specified. Additionally, some dyads were mixed gender while others were not. Although practically interesting, the current study would be underpowered in addressing these factors' impacts on cognitive workload. Previous research has also demonstrated that the emotional valence of conversation can influence various factors of performance (e.g., Hickman, Soccolich, Fitch, & Hanowski, 2015; McKnight, & McKnight, 1993; Nunes & Recarte, 2002). Although we could not control for the strength of this association, all dyads received the same discussion prompts and each member was given a turn to lead the discussion of the prompt.

The DRT method used herein to explore the dynamic ebb and flow of workload in a conversational dyad illustrates the potential for using this method for examining workload across two or more individuals working as a team in other operational environments. Based on the 2-DRT case used in the current research, having n-DRT units deployed across a team of individuals may provide insights into the flow of workload across the team as they perform a complex task. Future research should consider this possibility.

References

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

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American psychologist, 54*(7), 462.

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.

Boehm, U., Matzke, D., Gretton, M., Castro, S., Cooper, J., Skinner, M., ... & Heathcote, A. (2021). Real-time prediction of short-timescale fluctuations in cognitive workload. *Cognitive Research: Principles and Implications, 6*(1), 1-29.

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in cognitive sciences, 16*(2), 106-113.

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

Castro, S. C., Cooper, J. M., & Strayer, D. L. (2016). Validating two assessment strategies for visual and cognitive load in a simulated driving task. *Proceedings of the human factors and ergonomics society 56th annual meeting.* (Vol. 60, No. 1, pp. 1899-1903). Sage CA: Los Angeles, CA: SAGE Publications.

Castro, S. C., Heathcote, A., Cooper, J., & Strayer, D. (2021, November 23). Dynamic Workload Measurement and Modeling: Driving and Conversing. Retrieved from osf.io/4a9xb

Castro, S. C., Strayer, D. L., Matzke, D., & Heathcote, A. (2019). Cognitive workload measurement and modeling under divided attention. *Journal of Experimental Psychology: Human Perception and Performance, 45* 826-829.

Cooper, J., Castro, S.C., & Strayer, D.L., (2016). Extending the detection response task to simultaneously measure cognitive and visual task demands. In *Proceedings of the human factors and ergonomics society 56th 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.

Damaso, K. A., Castro, S. C., Todd, J., Strayer, D. L., Provost, A., Matzke, D., & Heathcote, A. (2021). A cognitive model of response omissions in distraction paradigms. *Memory & Cognition,* 1-17.

Drews, F. A., Pasupathi, M., & Strayer, D. L. (2008). Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied, 14*(4), 392-400. doi:10.1037/a0013119

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance: Vol. 2. Cognitive processes and performance* (41.1-41.49). New York: Wiley.

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., Lin, Y. S., Reynolds, A., Strickland, L., Gretton, M., & Matzke, D. (2019). Dynamic models of choice. Behavior research methods, 51(2), 961-985.

Hickman, J. S., Soccolich, S., Fitch, G., & Hanowski, R. J. (2015). Driver distraction: Eye glance analysis and conversation workload (No. FMCSA-RRR-14-001). United States. Federal Motor Carrier Safety Administration. Office of Analysis, Research, and Technology.

Howard, Z. L., Evans, N. J., Innes, R. J., Brown, S. D., & Eidels, A. (2020). How is multi-tasking different from increased difficulty?. *Psychonomic Bulletin & Review*, *27*(5), 937-951.

ISO DIS 17488 (2016). 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*. Englewood Cliffs, NJ: Prentice Hall.

Kubose, T. T., Bock, K., Dell, G. S., Garnsey, S. M., Kramer, A. F., & Mayhugh, J. (2006). The effects of speech production and speech comprehension on simulated driving performance. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *20*(1), 43-63.

Lee, A., Cerisano, S., Humphreys, K. R., & Watter, S. (2017). Talking is harder than listening: The time course of dual-task costs during naturalistic conversation. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *71*(2), 111.

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.

McKnight, A. J., & McKnight, A. S. (1993). The effect of cellular phone use upon driver attention. *Accident Analysis & Prevention, 25*(3), 259-265.

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

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

Norman, D. A., & Shallice, T. (1986). Attention to action. In *Consciousness and self-regulation* (pp. 1-18). Springer, Boston, MA.

Nunes, L., & Recarte, M. A. (2002). Cognitive demands of hands-free-phone conversation while driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *5*(2), 133-144.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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.

Regan, M., Hallett, C., & Gordon, C. P. (2011). *Accident Analysis and Prevention, 43,* 1771-1771.

Shinohara, K., Nakamura, T., Tatsuta, S., & Iba, Y. (2010). Detailed analysis of distraction induced by in-vehicle verbal interactions on visual search performance. *IATSS research*, *34*(1), 42-47.

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.

Strayer, D. L., Castro, S. C., & McDonnell, A. S. (In Press). The multitasking motorist. In A. Kiesel, L. Johannsen, H. Müller, & I. Hoch (Eds.) *The Handbook of Human Multitasking*, Springer.

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular phone. *Psychological Science*, 12, 462-466.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied*, 9, 23-52.

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J. R., Medeiros-Ward, N., & Biondi, F. (2015). Assessing cognitive distraction in the automobile. *Human Factors*, *57*(8), 1300-1324. doi:10.1177/0018720815575149

Teh, E., Jamson, S., Carsten, O., & Jamson, H. (2014). Temporal fluctuations in driving demand: The effect of traffic complexity on subjective measures of workload and driving performance. *Transportation research part F: traffic psychology and behaviour*, *22*, 207-217.

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.

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

Wickens, C. D., & McCarley, J. S. (2019). Applied attention theory. CRC press.

Author Note

The study design, hypotheses, and analytic plan for the current manuscript were not preregistered. The analytic code needed to reproduce the behavioral analyses is available at the following link: https://osf.io/4a9xb/ . Additionally, the modeling software necessary to fit the LBAO model to the data utilized in this study is also available within the LBAO_Modeling_ConvoDrive copy.zip file.

Appendix A

Conversation Question Prompts:

1. What was the worst school day you ever had?
2. What was the best school day you ever had?
3. What was the worst meal you ever cooked?
4. What was the best meal you ever cooked?
5. What was the worst outfit you ever wore?
6. What was the best outfit you ever wore?
7. What is the worst song that you love?
8. What is the song that you think is the best?
9. What is the worst photograph anyone ever took of you?
10. What is the best photograph anyone ever took of you?