Transportation Research Part F 26 (2014) 227–237

Contents lists available at ScienceDirect

Transportation Research Part F

journal homepage: www.elsevier.com/locate/trf

The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand

Ying Wang [a,b], Bryan Reimer [a], Jonathan Dobres [a,*], Bruce Mehler [a]

[a] MIT AgeLab, 77 Massachusetts Avenue, Cambridge, MA 02139, United States
[b] Beihang University, School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, 37 Xueyuan Road, Haidian District, Beijing 100191, China

ARTICLE INFO

Article history:
Received 27 August 2013
Received in revised form 29 May 2014
Accepted 4 August 2014

Keywords:
Driving
Cognitive workload
Gaze concentration
Eye movement
Eye tracking
Driver state

ABSTRACT

An observer's visual scanning behavior tends to narrow during periods of increased cognitive demand. Thus, measures of gaze concentration have become a popular method of gauging cognitive demand, but the consensus on the best method for computing gaze concentration is still evolving. This analysis considers measures of gaze concentration while driving an on-road vehicle, with and without two types of secondary cognitive demand (auditory and visuospatial working memory). We compare the advantages and disadvantages of several different methods for measuring gaze concentration, as well as a direct statistical comparison of their relative sensitivities. We find that several methods produce similar effect sizes, and that these are consistent across task types. Horizontal gaze dispersion, as measured from the standard deviation of horizontal gaze position, attained the largest effect size, indicating that it is the most sensitive to changes in gaze concentration under cognitive demand, while also being one of the simpler metrics to calculate. Our results show that complex eye tracking data sets from applied, ecologically valid situations such as on-road driving can be analyzed effectively with maximal sensitivity and minimal analytical burden to produce a robust measure of a driver's general allocation of attention.

1. Introduction

A diverse body of evidence suggests that cognitive processing is constrained by a variety of capacity limitations. These limitations have been demonstrated across a large number of experimental paradigms, including attentional blink (Raymond, Shapiro, & Arnell, 1992), auditory processing, (Kunar, Carter, Cohen, & Horowitz, 2008; May, Kennedy, Williams, Dunlap, & Brannan, 1990), working memory, (Fougnie & Marois, 2006), and visual search (Zelinsky, Rao, Hayhoe, & Ballard, 1997). When an observer's overall cognitive workload increases, his or her cognition and behavior may undergo compensatory changes that allow one task to be prioritized over another. For example, target processing in rapid serial visual presentation (RSVP) streams favors the first target (Raymond et al., 1992), eye movement patterns "zoom in" as a more detailed visual search is required (Zelinsky et al., 1997), and eye movement patterns contract as the difficulty of a tone counting task increases (May et al., 1990).

Cognitive-behavioral trade-offs under increased workload are also often observed outside the laboratory, particularly in the context of automotive research. The act of driving a motor vehicle, already a demanding task, has only grown more complicated

* Corresponding author. Tel.: +1 617 253 7728.
E-mail addresses: yingwang@buaa.edu.cn (Y. Wang), reimer@mit.edu (B. Reimer), jdobres@mit.edu (J. Dobres), bmehler@mit.edu (B. Mehler).

http://dx.doi.org/10.1016/j.trf.2014.08.003
1369-8478/

in recent years. Tasks undertaken while driving a vehicle have evolved from a set of relatively simple visual and manual demands, such as tuning the radio, to include an array of cognitively demanding activities, such as voice control of complex embedded vehicle systems and smartphone interactions. Visually demanding non-driving activities have a clear impact on a driver's ability to monitor the roadway. These activities can be characterized by measures such as total glance time, maximum glance time, etc. The Alliance of Automobile Manufacturers (2006), International Standards Organization (ISO, 2002) and more recently The National Highway Traffic Safety Administration (2013) provide guidance on the use of these metrics.

Increases in cognitive demand have been shown to impact drivers' allocation of attention to the roadway (Harbluk, Noy, Trbovich, & Eizenman, 2007; Liang & Lee, 2010; Reimer, Mehler, Wang, & Coughlin, 2010; 2012; Victor, Harbluk, & Engström, 2005). As cognitive demands increase, drivers are prone to concentrate their gaze directly in front of the vehicle. This gaze concentration is associated with a diminished ability to detect both peripheral and central targets, and a reduced frequency of glance to mirrors and the speedometer (Hammel, Fisher, & Pradhan, 2002; Harbluk et al., 2007; Liang & Lee, 2010; Nunes & Recarte, 2002; Recarte & Nunes, 2000; 2003; Strayer, Drews, & Johnston, 2003; Victor et al., 2005). These observations are consistent with a loss of situational awareness, inattentional blindness, and situations of ''look-but-fail-to-see'' (Kass, Cole, & Stanny, 2007; Recarte & Nunes, 2003; Reimer, Mehler, Wang, & Coughlin, 2012; Strayer et al., 2003). While various guiding principles have been established to characterize a glance-based assessment of visual demand, little consensus exists on how to most sensitively characterize changes in gaze concentration under added cognitive demand.

1.1. A brief history of gaze metrics

Eye gaze data have been used to characterize changes in cognitive load while driving for a number of years (Harbluk, Noy, & Eizenman, 2002; Sodhi et al., 2002; Victor et al., 2005). One of the more intuitive measures of changes in drivers' gaze under added cognitive demand is percent road center (PRC). PRC is defined as the percentage of fixations that fall within a predefined road center area during a specific period. PRC has been shown to increase with heightened cognitive demand (Engström, Johansson, & Östlund, 2005; Harbluk et al., 2002, 2007; Recarte & Nunes, 2000; Victor et al., 2005). While PRC is conceptually easy to understand, the definition of road center varies considerably across studies. Harbluk et al. (2007) employed a rectangular region 15° wide centered directly in front of the vehicle in the lane of travel. In one study presented in Victor et al. (2005), a rectangular region of 20° was considered. In another set of studies, Victor et al. (2005) defined the road center as a fixed circle with a diameter of 16° centered about the road center point. Extending this methodology further, Ahlstrom, Kircher, and Kircher (2009) considered a circular region, also 16° in diameter, centered on the driver's most frequent gaze angle.

Harbluk et al. (2007) based the computation of PRC on fixations identified from gaze positions. Other implementations of PRC (Ahlstrom et al., 2009; Victor et al., 2005) have used raw gaze points, a gaze trail recorded from an eye tracking system that is not clustered into fixations and saccades. In a comparison between these approaches, Ahlstrom et al. (2009) showed a strong correlation between fixation-based PRC and raw gaze-based PRC (mean $R = 0.95$).

The standard deviation of gaze points—whether computed from the observer's gaze angle or the projection of the gaze trail on a plane in front of the driver—have also been used to characterize changes in gaze behavior with cognitive demand (Reimer, 2009; Reimer et al., 2010, 2012; Sodhi et al., 2002; Victor et al., 2005). Sodhi et al. (2002) assessed the standard deviation of raw gaze points independently for the horizontal and vertical aspects of a driver's scan path. Reimer (2009), 2012) followed this conceptual model of independent analysis of the horizontal and vertical gaze components, while Victor et al. (2005) combined the vertical and horizontal components of gaze angle into a single gaze vector.

This paper examines the sensitivity of different approaches to characterizing changes in driver gaze in a field driving dataset encompassing periods of single task driving and driving under periods of moderate cognitive demand from two secondary tasks. The two approaches for computing PRC (fixation-based and unfiltered gaze-based), two methods of characterizing road center (rectangular world-based and circular gaze-based), and three approaches for assessing the standard deviation of gaze points (combined vertical and horizontal position, and the independent vertical and horizontal positions) are compared. Direct quantitative comparisons between the methods are presented to characterize their relative sensitivities.

2. Methods

2.1. Participants

This study uses data collected from the first session of a larger investigation of the driving behavior of older drivers (Dobres et al., 2013). The first session comprises baseline data for each subject, before any experimental interventions pertaining to the larger study were performed. A total of 38 subjects (21 males) between the ages of 60 and 75 took part in the study. Participants were required to be active, experienced drivers with a valid driver's license for at least the past 3 years. Participants were required to meet additional criteria: driving on average more than 3 times per week; having a driving record free of reported accidents for the past year; being in good health for their age; feeling comfortable driving a full-sized sedan; being free from major medical conditions or psychiatric disorders; and not currently taking a hypertensive prophylactic or somnolent drugs. Individuals who required corrective eyeglasses for normal driving were excluded in order to obtain robust eye tracking data (contact lenses were permitted). Recruitment was carried out in the greater Boston area. The study was approved by MIT's institutional review board.

One participant failed to adequately engage in the secondary tasks and was excluded from analysis. Two participants' eye tracking data were captured for less than 33% of the data analysis periods and were also excluded from the analysis. This left a total of 35 subjects (19 males) with a mean age of 67.7 years (SD = 5.18).

2.2. Apparatus

Data were collected in an instrumented 2010 Lincoln MKS sedan. Eye behavior was logged at 60 Hz using a faceLAB® 5.0 (Seeing Machines, Canberra, Australia) eye tracking system. Operation of the faceLAB system was carried out using "classic mode" with a manually created full head model.

2.3. Cognitive tasks

Drivers engaged in two secondary demand tasks during the driving session to increase cognitive workload (Fig. 1). The first was a verbal response delayed digit recall task. Subjects listened to a sequence of 10 auditorily presented digits, with the starting utterance of each digit spaced 2.25 s apart. Subjects had to repeat the digit in the sequence that was "one back" in the sequence from the currently presented digit. Each digit, 0–9, was presented once in the sequence in a randomized order. Participants responded to four such sequences over a 2-min demand period (Mehler, Reimer, & Coughlin, 2012; Reimer et al., 2012). Drivers also participated in a visual-spatial clock task (Paivio, 1978; Schlorholtz & Schieber, 2006). An auditory prompt presented a clock time, and subjects then had up to 10 s to verbally indicate whether the hands of the clock would form an acute angle (<90°) when set to the indicated time. Twelve clock prompts were presented over a 2-min period. While the 1-back primarily taxes the driver's auditory working memory, the clock task necessitates the visualization of the clock hands and taxes visuospatial working memory. Thus, changes in visual behavior can be assessed under two different types of secondary cognitive load.

2.4. Procedure

Participants were trained on the demand tasks to ensure proficiency prior to driving. Written instructions were provided, and participants were asked to respond to a practice trial (set of 10 stimuli per task) under the instruction of a research assistant. If participants missed more than 2 responses, additional practice trials were presented. Subsequently, participants reported to the instrumented vehicle and were instructed to adjust the seat and mirrors to their liking. The eye tracking system was then calibrated to create a full head model prior to driving. Prior to leaving the parking lot, participants again practiced the demand tasks and were provided with a safety briefing. During the course of the experiment, a research associate was seated in the rear of the vehicle to monitor data acquisition and driver safety. The research associate provided driving directions and answered essential questions during the experiment, while minimizing extraneous conversations.

The driving route consisted of approximately 10 min of urban driving to reach the interstate highway (I-93 N) and approximately 45 min of highway driving followed. Analysis periods of interest included a 2-min reference period of single task driving (baseline) before each demand period, a 2-min period of dual-task driving (driving while performing the cognitive tasks), and a 2-min single task recovery period. The baseline and recovery periods were separated from the demand periods by 30 s to ensure that they were in no way confounded by the secondary activities. The clock task was always administered on the outbound portion of the highway driving, while the 1-back task was always administered on the return portion of the drive. All instructions and task prompts were pre-recorded and played automatically during the driving session.

2.5. Eye tracker configuration

The faceLAB eye tracker is built around the concept of a "world model", a collection of virtual objects that mimics the configuration of the real-world recording environment, e.g. cockpit of the instrumented vehicle. This allows faceLAB to not only track the participant's head and eye positions, but also to estimate which objects the participant's gaze is focused

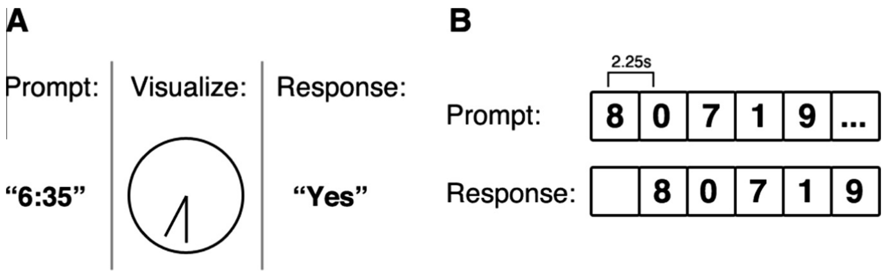

Fig. 1. Schematic illustrations of the two secondary tasks used during on-road driving periods. (A) The clock task, in which subjects were prompted with a time and were asked to visualize whether the hands of a clock set to that time would form an angle of less than 90°. (B) The 1-back task, during which subjects listened to sequences of digits presented 2.25 s apart and repeated the previous digit in the sequence. Although 5 digits are shown here for illustration, each sequence consisted of 10 digits.

on at any given time. An illustration of the world model as implemented for this experiment appears in Fig. 2. For measures of fixation-based and raw gaze-based PRC, the road center area was defined as a rectangle positioned directly in front of the driver (black rectangle in Fig. 2).

2.6. Processing of gaze points

The quality of the data collected by an eye tracking system in a field vehicle varies across a data set, owing to fluctuations in ambient light, vehicle motion, and the state of the participant's eyes (closure due to blinking, squinting, etc.). Artifact identification and quality control/filtering of the raw data are required to establish the reliability of data. Our group has explored a number of filtering rules in working with on-road driving datasets. The analyses in this paper were undertaken employing various combinations of the filters below, as defined in subsequent sections:

- *Condition 1*: the faceLAB automated gaze quality indices for the left and right eyes were reported as valid samples of eye position (gaze quality index $\geqslant 2$).
- *Condition 2*: faceLAB did not classify the sample as part of a blink.
- *Condition 3*: faceLAB did not classify the sample as part of a saccade.
- *Condition 4*: the sample gaze point intersected with a defined world model object.
- *Condition 5*: the estimated gaze point was within a bounding box subtending ±1.5 m horizontally and ±1.0 m vertically.
- *Condition 6*: the gaze point occurred within a group of at least 6 valid contiguous points (100 ms), defined as points that met the above 5 conditions (Salvucci & Goldberg, 2000).

2.7. Methods of calculating gaze concentration

2.7.1. Fixation-based PRC

Fixation-based PRC is the strictest method assessed for filtering the raw gaze data. In addition to meeting the six criteria listed above, a sample had to be part of a contiguous group of six samples that intersected the same world object, as fixation is targeted on a single region in space. Any sample that failed to meet all of these criteria was excluded from subsequent calculations. This approach to computing PRC is generally consistent with Harbluk et al. (2007). In Harbluk's study, fixations were calculated after the driving period based on data from a head-mounted eye tracker. In the present study, the intersection points of the driver's gaze with a world model object were computed in real-time by the faceLAB system. Fixation-based PRC was defined as the proportion of samples that fell within the road center rectangle in the faceLAB world model across all valid samples for a given experimental period (baseline, demand, and recovery).

2.7.2. Raw gaze-based ("unfiltered") PRC

Raw gaze measurement is a more general approach to looking at PRC that employs looser quality control criteria than in fixation-based PRC. The inclusion of this derived measure allows for the evaluation of how much of a difference a strict filter condition (fixation-based PRC) makes to the final PRC measurements. In the computation of this measure, we opted to include Condition 1 as the sole filter criterion. A gaze sample is considered valid as long as the faceLAB eye tracking system reported accurate tracking of both eyes. This greatly simplifies preprocessing at the expense of increasing the noise in the data (blinks, saccades, anomalous measurements, etc.). The road center area was defined as the same rectangle in the world model of the faceLAB system as in the fixation-based calculation. This analysis aims to provide a reference point more closely aligned with the approaches examined by Victor et al. (2005).

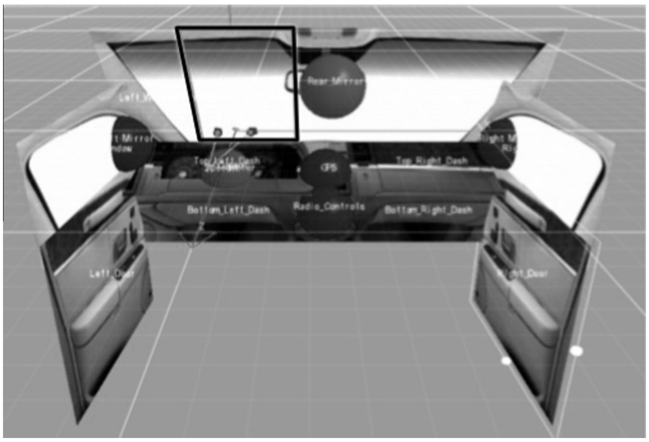

Fig. 2. Illustration of the faceLAB world model and road center object (black rectangle).

2.7.3. Central gaze-based PRC

Rather than defining the road center relative to world coordinates, this method defines it relative to the participant's most frequent gaze angle (Ahlstrom et al., 2009). The gaze angle data for all highway-driving portions of the session (including non-experimental measurement periods, spanning about twenty minutes prior to the any secondary demand) were binned into a 200 × 200 two-dimensional histogram subtending ±50° horizontally and vertically. The bin corresponding to the most frequent gaze angle was taken as the gaze center. If multiple angles were equally frequent, the one closest to a gaze angle of 0° was used. Road center was then defined as a circular area originating at the gaze center. A gaze plot for a representative participant is shown in Fig. 3. This procedure varied slightly from Ahlstrom et al. (2009) in that the histograms computed in the present study increase the resolution of the bins and consider a slightly narrower field of view. This approach was taken as glances beyond the 50° range are infrequent in this sample of highway driving.

Ahlstrom restricted the road center to a radius of 8°, but also suggested that other radii should be investigated. We therefore investigate the effects of defining the radius at 8°, 12°, and 16°. Because this definition relies on angular and not positional data, only Conditions 1, 2, and 3 were used to filter the samples.

2.7.4. Gaze dispersion

As described above, gaze *position* (not to be confused with gaze *angle*) refers to the physical point at which a participant's gaze intersects an object in the world model, measured in meters at a point near the windshield (position of the faceLAB cameras). Previous studies have shown that horizontal gaze dispersion is sensitive to the effects of cognitive demand (Reimer, 2009; Reimer et al., 2010; Sodhi et al., 2002). Gaze dispersion is defined as the standard deviation of participants' gaze positions and is measured separately for horizontal and vertical coordinates. Additionally, these two coordinates can be combined into a single measure of relative distance by applying the Pythagorean theorem ($a^2 + b^2 = c^2$) to them. The standard deviation of this measure may be thought of as "combined" gaze dispersion (Victor et al., 2005). A similar metric can also be computed based upon gaze angle data (Victor et al., 2005), which is largely equivalent to computations of gaze positions, though nominally less sensitive (see below). All six quality control conditions are employed to establish valid gaze points for gaze dispersion measures.

2.8. Statistical tests

The four gaze concentration measures (fixation-based PRC, unfiltered PRC, central gaze-based PRC, and gaze dispersion) were computed separately for three 2-min periods: baseline, cognitive demand (1-back or clock), and recovery (see *Cognitive Tasks* for details). This resulted in a 2 task × 3 period within-subject design for each measure. Analyses of the 1-back and clock demands are assessed independently to provide greater statistical clarity on the generalizability of different measures

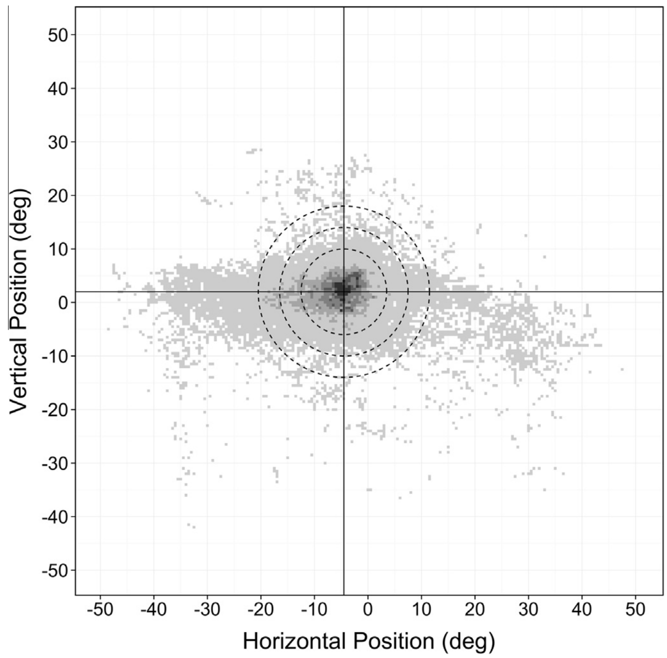

Fig. 3. Gaze angle histogram over 20 min of driving for a representative participant. Data have been binned and quantized for illustrative purposes, with darker colors indicating more frequent gazes. The crosshairs denote the driver's most frequent gaze angle, or "gaze center angle" (note that faceLAB's spatial origin point (0°) is relative to the position of the eye tracking cameras, which were mounted on the dashboard in line with the steering wheel, and not the road center). The dashed circles represent road center radii of 8°, 12°, and 16°.

of gaze concentration, and not whether the tasks produce differing levels of workload, *per se*. For a summary of workload considerations between these tasks see (Yang, Reimer, Mehler, & Dobres, 2013).

Careful consideration must be given to the type of statistical test used on data of this type. Percent road center is a ratio measurement (percentages summarize a numerator and denominator). Moreover, it is common to see road center percentages clustering near 100%, giving PRC distributions a strong negative skew. As a result, parametric statistics such as the t-test and repeated-measures ANOVA, which assume that the data follow an unskewed normal distribution, are inappropriate and likely to produce a Type I error. Instead, we employ Wilcoxon rank tests and the Friedman test (Friedman, 1937), which are the non-parametric analogs of the *t*-test and repeated-measures ANOVA, respectively. In contrast to measures of PRC, measures of gaze dispersion do not suffer these limitations, and appropriate parametric tests such as the *t*-test and repeated-measures ANOVA are used instead.

The purpose of the present research is to compare the sensitivity of several measures of gaze concentration. Therefore, it would be useful to gauge the effect sizes of each measurement method irrespective of scale range. Effect size is a method of quantifying the size of a difference, as opposed to assessing whether a difference is statistically significant. Common effect size measures such as Cohen's *d* and partial eta-squared rely on the means and standard deviations of the samples being compared, and assume that samples are unskewed and normally distributed. Cliff's delta provides a suitable non-parametric measure of effect size (Cliff, 1993). Cliff's delta makes no assumptions about the data's underlying distribution and directly compares the degree of overlap between two groups of observations. For an excellent illustration of how Cliff's delta is computed and how the effect size categories are derived, see (Romano, Kromrey, & Coraggio, 2006).

Gaze concentration metrics were computed for periods during and surrounding a visuospatial clock task and an auditory n-back task. This was done primarily to ensure that the chosen metrics were robust across different types of demand, but not to compare the demands of the two tasks, *per se*. A full examination of similarities and differences between these tasks is beyond the scope and intentions of the present research, as there are numerous theoretical differences in the cognitive underpinnings of the tasks Therefore, task order was not counterbalanced.

3. Results

3.1. Cognitive task performance

Across all the experimental trials, participants responded to the 1-back task with a mean accuracy of 94.5% (SD = 7.2%), and to the clock task with a mean accuracy of 88.5% (SD 14.1%). The high accuracy on these tasks suggests that participants were actively engaged.

3.2. World-based PRC (fixation and unfiltered)

As shown in Fig. 4, participants' gaze patterns contracted toward the road center while performing the digit recall and clock tasks, resulting in higher PRCs in the demand period as compared to baseline and recovery periods. The stringently filtered fixation-based PRC measurements show a significant difference between task periods for the 1-back task ($X^2_{(df = 2)} = 14.82$, $p < .001$, Friedman test) as well as the clock task ($X^2_{(df = 2)} = 8.69$, $p < .013$). During the 1-back task, PRC is significantly elevated during the demand period relative to the other periods (Demand vs. Recovery $p < .001$, Demand vs. Baseline $p = .002$, Baseline vs. Recovery $p = .802$, Wilcoxon signed rank tests). A similar pattern is evident during the clock task (Demand vs. Recovery $p = .025$, Demand vs. Baseline $p = .054$, Baseline vs. Recovery $p = .270$).

The same pattern of results is seen for the unfiltered PRC measure. PRC is elevated during the 1-back demand period ($X^2_{(df = 2)} = 17.65$, $p < .0001$ for period, Friedman test; Demand vs. Recovery $p < .001$, Demand vs. Baseline $p = .001$, Baseline vs. Recovery $p = .481$, Wilcoxon signed rank tests), and trends toward significance for the clock task period ($X^2_{(df = 2)} = 5.20$, $p = .074$) A comparison between the fixation-based and unfiltered measurement methods shows that overall PRC is slightly lower for the unfiltered measure (84.7% for unfiltered vs. 87.6% for fixation-based measures). This difference is small but consistent, resulting in a significant difference between methods ($p < .001$ for both demand types, Wilcoxon signed rank test).

3.3. Central gaze-based PRC

Central gaze-based PRC measures are shown in Fig. 5. When the road center area is defined as a circle with radius 8° centered on the participant's most frequent gaze angle, no significant changes in PRC are observed across task periods (1-back task, $X^2_{(df = 2)} = 1.14$, $p = .567$; clock task, $X^2_{(df = 2)} = 2.32$, $p = .313$). Central gaze-based PRC averages 67.9% (1-back) and 69.9% (clock) across the three measurement periods, lower than both fixation-based PRC (87.2% during 1-back, 87.9% during clock) and unfiltered PRC (85.5% during 1-back, 83.9% during clock). PRC is significantly lower when measured with this gaze-based method ($p < .0001$ when comparing central gaze-based PRC to fixation-based and unfiltered PRC for both tasks, Wilcoxon signed rank test).

When the radius of the road center area is expanded to 12°, overall central gaze-based PRC increases to 82.6% (1-back) and 84.4% (clock), which is slightly lower than the world-based measures (1-back, $p = .031$ and $p = .254$, Wilcoxon signed rank test between central gaze-based PRC and fixation-based and unfiltered PRC, respectively; clock, $p = .106$ and

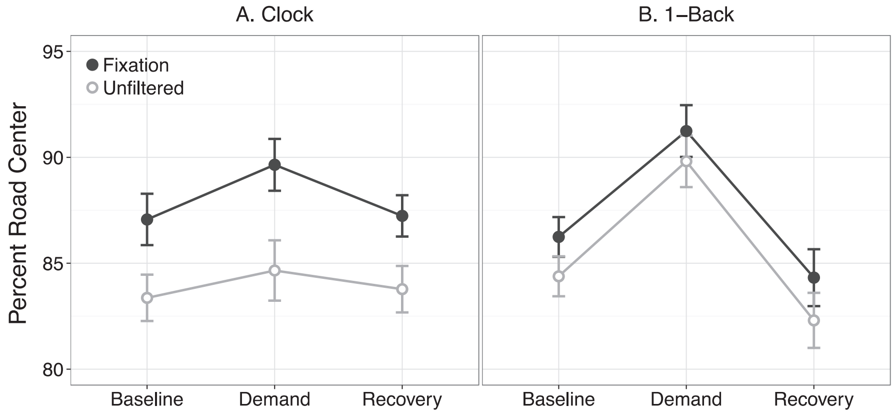

Fig. 4. Fixation and unfiltered measures of PRC during baseline, demand, and recovery periods for (A) the clock visualization task and (B) the 1-back task. Error bars represent ± 1 SEM (within-subject).

$p = .522$). During the 1-back task, PRC increases significantly during the demand period compared to baseline and then decreases slightly ($X^2_{(df=2)} = 8.17$, $p = .017$, Friedman test on period; Baseline vs. Demand $p = .009$, Recovery vs. Demand $p = .107$, Baseline vs. Recovery $p = .788$, Wilcoxon signed rank tests). Similarly, during the clock task, central gaze-based PRC is significantly elevated ($X^2_{(df=2)} = 6.88$, $p = .032$, Friedman test on period; Baseline vs. Demand $p = .024$, Recovery vs. Demand $p = .024$, Baseline vs. Recovery $p = .787$, Wilcoxon signed rank tests).

When the radius of the road center area is expanded to 16°, overall central gaze-based PRC increases to 88.9% (1-back) and 90.8% (clock), roughly comparable to world-based PRC measures (1-back, $p = .222$ and $p = .015$, Wilcoxon signed rank test between central gaze-based PRC and fixation-based and unfiltered PRC, respectively; clock, $p = .018$ and $p = .007$). During the 1-back task, central gaze-based PRC increases significantly during the demand period compared to baseline and then decreases significantly ($X^2_{(df=2)} = 11.17$, $p = .002$, Friedman test on period; Baseline vs. Demand $p = .007$, Recovery vs. Demand $p = .02$, Baseline vs. Recovery $p = .754$, Wilcoxon signed rank tests). Similarly, during the clock task, central gaze-based PRC increases significantly during the demand period compared to baseline and then decreases significantly ($X^2_{(df=2)} = 14.53$, $p < .001$, Friedman test on period; Baseline vs. Demand $p = .004$, Recovery vs. Demand $p < .0001$, Baseline vs. Recovery $p = .091$, Wilcoxon signed rank tests).

This analysis implies that at a small circular aperture representing the road center, the percentage of time a driver's gaze intersects with this region is insensitive to changes in cognitive demand. Furthermore, defining a conservative road center area will mask changes in PRC. A broader 12° or 16° radius road center area provides increased sensitivity to changes in cognitive demand. Interestingly, a more consistent pattern of PRC increase appears between the two tasks using the gaze-based PRC metrics as compared to the fixation-based and unfiltered PRC metrics.

3.4. Gaze dispersion

Fig. 6 illustrates changes in gaze dispersion across the three measurement periods. During the 1-back task, combined gaze dispersion decreases significantly during the demand period and rebounds during the recovery period ($F_{(2, 68)} = 8.25$, $p < .001$,

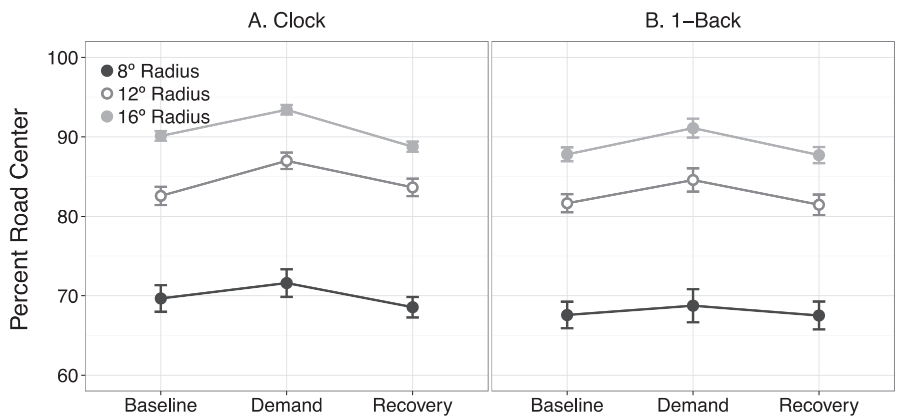

Fig. 5. PRC measurements during the three task periods when road center is defined as a circular area originating from the driver's most frequent gaze angle. (A) clock task, (B) 1-back task.

repeated-measures ANOVA; Baseline vs. Demand $p = .01$, Demand vs. Recovery $p = .02$, Baseline vs. Recovery $p = .97$, paired *t*-tests). Horizontal gaze dispersion decreases significantly during the demand period and rebounds during the recovery period ($F_{(2, 68)} = 15.1$, $p < .0001$, repeated-measures ANOVA; Baseline vs. Demand $p = .001$, Demand vs. Recovery $p = .003$, Baseline vs. Recovery $p = .908$, paired *t*-tests). Vertical gaze dispersion does not change significantly between periods ($F_{(2, 68)} = 0.11$, $p = .895$, repeated-measures ANOVA).

Results are similar during the clock task. Combined gaze dispersion decreases significantly during the demand period and rebounds during the recovery period ($F_{(2, 68)} = 10.39$, $p < .001$, repeated-measures ANOVA; Baseline vs. Demand $p = .032$, Demand vs. Recovery $p = .012$, Baseline vs. Recovery $p = .839$, paired *t*-tests). Horizontal gaze dispersion decreases significantly during the task period and rebounds during the recovery period ($F_{(2, 68)} = 24.5$, $p < .0001$, repeated-measures ANOVA; Baseline vs. Task $p = .002$, Task vs. Recovery $p < .001$, Baseline vs. Recovery $p = .738$, paired t-tests). Vertical gaze dispersion does not change significantly between periods ($F_{(2, 68)} = 1.37$, $p = .260$, repeated-measures ANOVA).

Consistent with previous research, these results indicate that horizontal gaze dispersion is highly sensitive to changes in cognitive demand, while vertical gaze dispersion is not (Reimer, 2009; Reimer et al., 2010; Sodhi et al., 2002). Combined gaze dispersion is a spatial average of these two component measures and is therefore less affected by cognitive demand than the horizontal measure alone. However, as in Victor et al. (2005), the combined measure changes significantly under cognitive demand. Consistent with the central gaze-based PRC measures, there is overall consistency of patterns in gaze dispersion between the two tasks.

3.5. Effect size

Table 1 presents Cliff's Delta values comparing the baseline and demand periods for each measurement method (absolute values are shown). Within demand type, effect sizes for most measures are similar and consistent with the plots in Figs. 4–6. Effect sizes are noticeably smaller for the clock task as compared to the 1-back. For both types of demand, horizontal gaze dispersion has the largest associated Cliff's Delta value. Within the clock task, this is the only measurement method to reach a medium effect size, and for the 1-back, this nearly qualifies as a "large" effect size ($\geqslant 0.474$, see Romano et al. (2006) for categorical derivations). This suggests that among all the methods and tasks assessed, horizontal gaze dispersion is the most sensitive to changes in gaze patterns under added cognitive demand. To investigate potential differences between the use of gaze position and gaze angle data for the calculation of gaze dispersion, we also calculated effect sizes for horizontal gaze dispersion measures based on gaze angle. We found that the effect sizes were similar to those of gaze position (.433 for the 1-back task, .421 for the clock task). Though the gaze angle effect sizes are nominally smaller as compared to the dispersion measures based on gaze position, they are still substantially larger than other measures of gaze concentration. Since dispersion metrics based on gaze angle were statistically equivalent to those based on gaze position, we opt not to report those results in detail for the sake of clarity.

4. Discussion

This analysis shows that raw gaze data can be used to accurately track changes in gaze concentration (Ahlstrom et al., 2009; Reimer et al., 2010; Sodhi et al., 2002; Victor et al., 2005). As initially proposed by Sodhi et al. (2002), further demonstrated by Victor et al. (2005), and elaborated by Ahlstrom et al. (2009), and reported on in Reimer et al. (2012), raw gaze data do not need to be segmented into a pattern of fixations before calculating changes in gaze behavior. Simpler filtering criteria adequately reduce noise in the data. Moreover, data analysis is not constrained by the fitting of subjectively defined regions (fixations) to the data. Ahlstrom et al. (2009) showed that PRC measures calculated with and without fixation

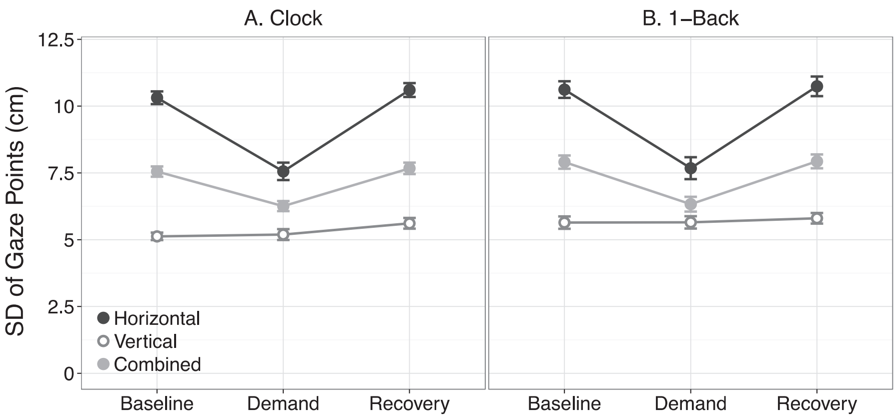

Fig. 6. Standard deviation of horizontal, vertical, and combined gaze position across the three measurement periods for (A) clock task and (B) 1-back task.

Table 1
Effect sizes for each measurement type. Cells marked "NA" were not computed because they failed to show statistically significant differences between task periods.

Gaze measure	1-Back		Clock	
World-based fixation	0.346	Medium	0.201	Small
World-based unfiltered	0.351	Medium	NA	NA
Gaze-based 8°	NA	NA	NA	NA
Gaze-based 12°	0.239	Small	.249	Small
Gaze-based 16°	0.332	Medium	.265	Small
Gaze dispersion Horizontal	0.463	Medium	.442	Medium
Gaze dispersion vertical	NA	NA	NA	NA
Gaze dispersion combined	0.352	Medium	.329	Small

segmentation, corresponding to our "fixation" and "unfiltered" methodologies, are highly correlated. The results of the present study show that effect sizes (Cliff's Delta) from the highly filtered fixation-based PRC measurements and the unfiltered PRC measurements are consistent. The only notable difference between these measures is a decrease in the magnitude of unfiltered PRC relative to fixation based PRC. This suggests that a larger percentage of the unfiltered samples fall outside of the road center area. This observation may, in part, be related to a reduction in the reliability of eye positions as the eyes move in directions divergent from the roadway and out of view of the eye tracking system's primary optics.

Ahlstrom et al. (2009) reported that unfiltered PRC measures are somewhat larger than fixation-based PRC measures, the opposite of the pattern we observe. It is worth bearing in mind that Ahlstrom's study reported a much lower overall PRC (a mean of 36% for fixation-based data and 44% for unfiltered data) than what had previously been found for attentive drivers in the field (approximately 70%, Victor et al., 2005) and 67–91% based upon the different methods presented here. Differences in reported PRC may relate to the environment in which the data were collected. While Ahlstrom et al. (2009) utilized local roadways, this study, as well as Victor et al. (2005), was conducted on a highway. In highway travel, off-road-center fixations would be expected to occur less frequently than in local driving situations. Environmental and algorithmic factors may impact the overall magnitude of PRC measures, but the present study suggests that PRC magnitude changes are independent of sensitivity, consistent with previous work.

In the original conception of fixation-based PRC measures (Harbluk et al., 2002), the road center was defined as a rectangle 15° wide. Victor et al. (2005) modified this to a circle with a diameter of 16° centered at the road center. This was later broadened by Ahlstrom et al. (2009) to consider a region around the driver's most frequent gaze angle. The first two methods position the road center area relative to a defined region of the roadway. In contrast, central gaze-based PRC defines the location of the road center based on the pattern of drivers' gaze points. This provides a more flexible method of accounting for differences in behavior between individual drivers. In this study, the radius of the road center circle was initially set to 8° (16° in diameter). Although this resulted in PRC measurements that were not statistically impacted by changes in cognitive demand, this is in line with previous field research (Victor et al., 2005). In the field driving conditions, Victor et al. (2005) reported PRC in the 70% range. This is highly consistent with the 67% reported here. Enlarging the road center area to a radius of 12° or 16°, however, increased the sensitivity of PRC to changes in cognitive demand. Our highway data suggests that shorter radii, such as 8°, are not optimal in the computation of PRC.

It is worth emphasizing that the world-based measures of PRC (both filtered into fixations and unfiltered) had smaller effect sizes and were less sensitive to demand-induced changes in visual behavior than measures of gaze-based PRC and gaze dispersion. Moreover, the patterns of PRC increase and decrease across measurement periods were inconsistent between tasks when measured with world-based methods.

The process of identifying fixations in large datasets is tedious and time consuming, and is made more so in dynamic environments, as in as driving, where the driver's head movements, smooth eye movements, lighting conditions, and other sources of data recording errors impair tracking quality and make decisions about the treatment of missing data subjective (Reimer & Sodhi, 2006). Furthermore, characterizing a driver's most frequent gaze angle is computationally demanding. In contrast, the horizontal gaze dispersion measure provides better discrimination of demand through a methodology that is also considerably simpler. The enhanced sensitivity of the horizontal gaze dispersion measure over the vertical or combined measure is likely related to the ecological demands of the driving task (Reimer et al., 2012) as well as the biological constraints of human eye movement (Steinman, Haddad, Skavensk, & Wyman, 1973). Specifically, drivers' scan paths are characterized by frequent orientation toward the roadway environment and less frequent glances toward the dashboard and other vertically displaced objects. The sensitivity of this measure remains regardless of whether gaze position or gaze angle data are used.

This work shows that there are a variety of reasonably sensitive ways to measure changes in gaze concentration that occur under increased cognitive demand. The effect of a cognitive task on gaze concentration was statistically significant for all measures except central gaze-based PRC with an 8° road center radius and vertical gaze dispersion across both tasks, and PRC measures for the clock task based on unfiltered gaze data. Among the sensitive measures, several had effect sizes in a relatively narrow range (around 0.35, see Table 1). These included fixation-based PRC, unfiltered PRC, combined gaze dispersion, and gaze-based PRC at a radius of 16°. Gaze-based PRC at a radius of 12° had a smaller effect size, suggesting that it is considerably less sensitive than these other measures. Horizontal gaze dispersion, on the other hand, had a substantively larger effect size, indicating that horizontal gaze dispersion has the highest sensitivity to changes in gaze concentration under task demand.

While reports using horizontal gaze dispersion to characterize changes in cognitive workload have appeared previously (Reimer, 2009; Reimer et al., 2010), other reports have focused on less sensitive metrics that combine the horizontal and vertical aspects of gaze (Victor et al., 2005). The present data make it clear that, as discussed in Reimer et al. (2012), merging the horizontal and vertical gaze dispersion components does not produce a more sensitive measure. If anything, it mutes the sensitive horizontal dispersion measure by spatially averaging it with the uninformative vertical dispersion measure.

Gaze data are voluminous and complex. PRC measures are attractive because they simplify the translation of gaze data into a measure of visual attention. Regardless of the specific implementation, data points are ultimately sorted into "road center" and "not road center" groups. In contrast to this dichotomous approach, gaze dispersion is a continuous measure, and thus is able to summarize the underlying gaze data with a greater degree of granularity and sensitivity. It is also worth noting that gaze dispersion does not suffer the "bounding problem" that PRC measures do, in which PRC measurements may cluster at the high end of the measurement scale. Moreover, gaze dispersion is straightforward to calculate and does not rely on the creation of an arbitrary road center area. These characteristics make horizontal gaze dispersion an efficient and highly sensitive measure of visual behavior.

Recent efforts to create real-time driver state monitoring systems have shown that eye behavior data is approximately as sensitive to changes in driver state as vehicle telemetry data, and that when used in combination with other metrics, eye data substantially increases the sensitivity of state detection systems. (Solovey, Zec, Perez, Reimer, & Mehler, 2014; Tan, Reimer, Mehler, & Coughlin, 2011). Although the present research was not conducted with driver state monitoring in mind, it stands to reason that having a simple, robust method for characterizing gaze behavior in real-time may prove to be a valuable addition to these types of systems.

However, it should be kept in mind that the sensitivity of the dispersion metric is dependent on the resolution of the eye-tracker. If the bins are too coarse or sparse, or are of unequal size (perhaps due to errors during data collection or environmental limitations), inaccuracies may be introduced into the calculation. Measures of PRC, though less sensitive, do not suffer from this limitation. Finally, it is important to note that the use of dispersion-based measures is limited to comparing segments of data gathered from time periods of relatively consistent length. In essence, the number of sample points included in a time period impacts the degree of "noise" in the computation of variability. Since PRC measures reduce glance samples into a single ratio measurement, they are less susceptible to this methodological obstacle.

Lastly, it is worth bearing in mind that this study examined a sample of older drivers, which may raise concerns as to the study's generalizability. Although behavioral and cognitive differences between younger and older drivers should not be ignored in on-road studies, prior research suggests that the types of changes in gaze concentration observed in the present study do not differ significantly across age groups (Reimer et al., 2010; Yang et al., 2013). The tasks employed in the current study tended to draw drivers' eyes toward the forward roadway, which may have led to a uniformity of gaze behavior across subjects, thus rendering age differences inconsequential in the methodological comparison of measures. Contexts that might call upon different cognitive resources may reveal more age-related differences in gaze behavior, and present a promising avenue for future study.

5. Conclusions

This study examined the sensitivity of a variety of eye movement measures that have been used to characterize gaze centralization in response to cognitive demands. The results demonstrate that horizontal gaze dispersion is the most sensitive measure of cognitive demand in these situations. PRC measurements based on unfiltered or "noisy" eye gaze data also showed good sensitivity, indicating that it is unnecessary to segment raw gaze data into fixation patterns before analysis. Regardless of the method selected, the centralization of gaze observed under increased cognitive workload appears to occur across a wide variety of demands reported in the literature. However, the degree to which this reduction in scan path impacts awareness of the roadway, and how these measures correspond to impairments in driving safety, remain unknown. Finally, it is worth noting that the methodological comparison presented here is based upon data gathered from a single eye tracking system over a fixed time window. The world-based and unfiltered PRC methods rely heavily on the eye tracking system's automated assignment of gaze to an object area, while the gaze-based measures do not. Future work will need to establish the validity of the automated processing algorithms within the faceLAB system and the degree to which these algorithms impact the reliability of these comparisons, as well as how these results generalize to other eye tracking systems.

Acknowledgments

Support for this research was provided by the United States Department of Transportation's Region One, New England University Transportation Center and the Santos Family Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.trf.2014.08.003.

References

Ahlstrom, C., Kircher, K., & Kircher, A. (2009). Considerations when calculating percent road centre from eye movement data in driver distraction monitoring. In: 5th International driving symposium on human factors in driver assessment, Big Sky, MT, US, 132–139.

Alliance of Automobile Manufacturers. (2006). Statement of principles, criteria and verification procedures on driver interactions with advanced in-vehicle information and communication systems, pp. 1–90.

Cliff, N. (1993). Dominance statistics – ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*(3), 494–509.

Dobres, J., Potter, A., Reimer, B., Mehler, B., Mehler, A., & Coughlin, J. (2013). Assessing the impact of "brain training" on changes in driving performance, visual behavior, and neurophysiology. Driving Assessment Conference: 8th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design.

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour, 8*(2), 97–120. http://dx.doi.org/10.1016/j.trf.2005.04.012.

Fougnie, D., & Marois, R. (2006). Distinct capacity limits for attention and working memory: Evidence from attentive tracking and visual working memory paradigms. *Psychological Science, 17*(6), 526–534. http://dx.doi.org/10.1111/j.1467-9280.2006.01739.x.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701.

Hammel, K. R., Fisher, D. L., & Pradhan, A. K. (2002). Verbal and spatial loading effects on eye movements in driving simulators: A comparison to real world driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 46*(26), 2174–2178. http://dx.doi.org/10.1177/154193120204602615.

Harbluk, J. L., Noy, Y. I., & Eizenman, M. (2002). The impact of cognitive distraction on driver visual behaviour and vehicle control. Transport, Canada.

Harbluk, J. L., Noy, Y. I., Trbovich, P. L., & Eizenman, M. (2007). An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. *Accident Analysis and Prevention, 39*(2), 372–379. http://dx.doi.org/10.1016/j.aap.2006.08.013.

ISO 15007–1:2002. Road vehicles – measurement of driver visual behaviour with respect to transport information and control systems – part 1: Definitions and parameters. TC22/SC13, International Organization for Standardization.

Kass, S. J., Cole, K. S., & Stanny, C. J. (2007). Effects of distraction and experience on situation awareness and simulated driving. *Transportation Research Part F: Traffic Psychology and Behaviour, 10*(4), 321–329. http://dx.doi.org/10.1016/j.trf.2006.12.002.

Kunar, M. A., Carter, R., Cohen, M., & Horowitz, T. S. (2008). Telephone conversation impairs sustained visual attention via a central bottleneck. *Psychonomic Bulletin and Review, 15*(6), 1135–1140. http://dx.doi.org/10.3758/PBR.15.6.1135.

Liang, Y., & Lee, J. D. (2010). Combining cognitive and visual distraction: Less than the sum of its parts. *Accident Analysis and Prevention, 42*(3), 879–888. http://dx.doi.org/10.1016/j.aap.2009.05.001.

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta Psychologica, 75*(1), 75–89.

Mehler, B., Reimer, B., & Coughlin, J. F. (2012). Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: An on-road study across three age groups. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 54*(3), 396–412.

National Highway Traffic Safety Administration (2013). Visual–manual NHTSA driver distraction guidelines for in-vehicle electronic devices. NHTSA-2010-0053, pp. 1–281.

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. http://dx.doi.org/10.1016/S1369-8478(02)00012-8.

Paivio, A. (1978). Comparisons of mental clocks. *Journal of Experimental Psychology: Human Perception and Performance, 4*(1), 61–71.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance, 18*(3), 849–860.

Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied, 6*(1), 31–43. http://dx.doi.org/10.1037/1076-898X.6.1.31.

Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: Effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology: Applied, 9*(2), 119–137. http://dx.doi.org/10.1037/1076-898X.9.2.119.

Reimer, B. (2009). Impact of cognitive task complexity on drivers' visual tunneling. *Transportation Research Record: Journal of the Transportation Research Board, 2138*, 13–19.

Reimer, B., Mehler, B., Wang, Y., & Coughlin, J. F. (2010). The impact of systematic variation of cognitive demand on drivers' visual attention across multiple age groups. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54*(24), 2052–2055. http://dx.doi.org/10.1177/154193121005402407.

Reimer, B., Mehler, B., Wang, Y., & Coughlin, J. F. (2012). A field study on the impact of variations in short-term memory demands on drivers' visual attention and driving performance across three age groups. *Human Factors, 54*(3), 454–468. http://dx.doi.org/10.1177/0018720812437274.

Reimer, B., & Sodhi, M. (2006). Detecting eye movements in dynamic environments. *Behavior Research Methods, 38*(4), 667–682.

Romano, J., Kromrey, J., & Coraggio, J. (2006). Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys. In: Presented at the Annual Meeting of the Florida Association of Institutional Research, Cocoa Beach, FL.

Salvucci, D., & Goldberg, J. (2000). Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the eye tracking research and applications symposium, pp. 71–78.

Schlorholtz, B., & Schieber, F. (2006). Assessment of age differences in mental workload while driving using verbal versus visual-spatial subsidiary tasks (Vol. 50, pp. 2378–2382). In: Presented at the proceedings of the human factors and ergonomics society's 50th annual meeting, Sage Publications.

Sodhi, M., Reimer, B., Cohen, J. L., Vastenburg, E., Kaars, R., & Kirschenbaum, S. (2002). On-road driver eye movement tracking using head-mounted devices. In: Proceedings of the 2002 symposium on eye tracking research & applications, 61–68.

Solovey, E., Zec, M., Perez, E. A. G., Reimer, B., & Mehler, B. (2014). Classifying driver workload using physiological and driving performance data: Two field studies. In: Presented at the CHI conference on human factors in computing systems, Toronto, Canada, pp. 1–10.

Steinman, R. M., Haddad, G. M., Skavensk, A. A., & Wyman, D. (1973). Miniature eye-movement. *Science, 181*(4102), 810–819.

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*(1), 23–32.

Tan, Z., Reimer, B., Mehler, B., & Coughlin, J. F. (2011). Detection of elevated states of cognitive demand in a naturalistic driving environment. In: Presented at the 2nd annual international conference on advanced topics in artificial intelligence, pp. 1–6.

Victor, T. W., Harbluk, J. L., & Engström, J. A. (2005). Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transportation Research Part F: Traffic Psychology and Behaviour, 8*(2), 167–190. http://dx.doi.org/10.1016/j.trf.2005.04.014.

Yang, Y., Reimer, B., Mehler, B., & Dobres, J. (2013). A field study assessing driving performance, visual attention, heart rate and subjective ratings in response to two types of cognitive workload. In: Proceedings of the 7th international driving symposium on human factors in driver assessment, training, and vehicle design. Bolton Landing, NY, pp. 397–403.

Zelinsky, G. J., Rao, R. P., Hayhoe, M. M., & Ballard, D. H. (1997). Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science*, 448–453.