Sensitivity of Multiple Cognitive Workload Measures: A Field Study Considering Environmental Factors

Joonwoo Son*
HumanLAB
Daegu Gyeongbuk Institute of Science & Technology
Deagu, South Korea
+82 53 785 4740
json@dgist.ac.kr

Myoungouk Park
HumanLAB
Daegu Gyeongbuk Institute of Science & Technology
Deagu, South Korea
+82 53 785 4741
violet1211@dgist.ac.kr

Hosang Oh
HumanLAB
Daegu Gyeongbuk Institute of Science & Technology
Deagu, South Korea
+82 53 785 4754
Ohs2384@dgist.ac.kr

ABSTRACT

Objective: This paper aims to compare the sensitivity of multimodal cognitive workload measures for classifying a driver's cognitive demand level from on-road experimental data. The measurement domains consist of driving performance, physiological arousal and eye behavioral change. **Method:** subjects (15 males in the 25-35 age range (M=27.9, SD=3.13)), experimental setup (an instrumented vehicle which consists of six video cameras, driving data logger, gaze tracker, and physiological measurement systems), procedure (20 minutes of driving exercise on a urban road and another 20 minutes of highway driving on 36km of highway), cognitive load (N-back task, an auditory delayed digit recall task was used to create periods of cognitive demand at three distinct levels), rating of driving workload (rating subjective driving workload after watching the experimental video clips by 4 different reviewers). **Result:** Potential measures of driver's cognitive workload are suggested to estimate drivers' cognitive demand level. It is expected that these measures can be used for evaluating the design of in-vehicle interface objectively.

Categories and Subject Descriptors

J.4 [Social and Behavioral Sciences]

General Terms

Human Factors

Keywords

Cognitive Workload Metrics Sensitivity, Driving Performance, Physiology, Eye Behavior, Driving Workload

1. INTRODUCTION

In recent years, a large number of researchers have been devoted to investigating the effect of cognitive workload on driving performance [1-2], physiological response [3-4], and eye behavior [5-6]. However, it is know that there is no simple measure to index cognitive workload because the driver's mental status is not observable, and each of these methods provides advantages and disadvantages depending on the setting and measurement goal. Among those measurement domains, driving performance measures can detect the cognitive workload using easy and less expensive methods, but have limitations compared to others due to small changes according to the cognitive workload [2]. For the physiological measures, several driving research projects were examining physiological measures as indicators of workload during driving. Most recently, Mehler et al. found that both heart rate and skin conductance were sensitive physiological measures for detecting systematic variations in cognitive demand [3]. These findings are different from the results of the HASTE project that show inconsistent relationships between heart period, skin conductance, and demand level in both auditory and visual tasks and do not suggest any consistently sensitive physiological measures for differentiating cognitive demand levels [4]. In the eye behavioral measures, Reimer et al. reported that horizontal gaze concentration under systematically added cognitive demand, which was loaded by the same surrogate secondary task as that of Mehler's study, increases in a relatively linear fashion [5]. However, the results of the HASTE project suggested that significant gaze concentration caused by the auditory task in comparison with baseline but they do not show significant increase in gaze concentration between tasks [6].

In order to clarify those conflicted findings, this field study aims for replicating systematically added cognitive demand method in a different setting and comparing the sensitivity of multiple cognitive measures for differentiating four levels of cognitive demand from a working memory task. In addition, this study evaluates the primary driving workload rates [7] to consider the influence of environmental factors when comparing these results with other field studies.

2. METHOD

2.1 Field Study with Cognitive Load

2.1.1 Subject

Subjects were required to meet the following criteria: age between 25-35, drive on average more than twice a week, be in self-reported good health and free from major medical conditions, not take medications for psychiatric disorders, score 25 or greater on the mini mental status exam to establish reasonable cognitive capacity and situational awareness. The subjects consisted of 15 young males (M=27.9, SD=3.13).

2.1.2 Experimental setup

The experiments were conducted in a full size sedan that is instrumented for collecting time-synchronized data. The DGIST instrumented vehicle consists of six video cameras (two for driver and four for road environment monitoring), high speed and low speed CAN logger, driver gaze tracking system, and physiological measurement system. The DGIST-designed custom monitoring software was separately running on four windows-based PCs and synchronized by storing the measurement data with master time that was sent by a main control PC.

Copyright held by author(s)
AutomotiveUI'12, October 17-19, Portsmouth, NH, USA.
Adjunct Proceedings

2.1.3 Cognitive workload

An auditory delayed digit recall task, so called n-back task was selected to create periods of cognitive demand at three different levels. This form of n-back task requires participants to say out loud the nth stimulus back in a sequence that is presented via audio recording [5]. The lowest level n-back task is the 0-back where the participant is to immediately repeat out loud the last item presented. At the moderate level (1-back), the next-to-last stimulus is to be repeated. At the most difficult level (2-back), the second-to-the-last stimulus is to be repeated. The n-back was administered as a series of 30-second trials consisting of 10 single digit numbers (0-9) presented in a randomized order at an inter-stimulus interval of 2.1 seconds. Each task period consisted of a set of four trials at a defined level of difficulty resulting in demand periods that were each two minutes long. The n-back task was pre-trained until the participants met minimum performance criteria (No error for 0-back, not more than two errors for 1-back, and not more than three errors for 2-back). This n-back task procedure replicated the method of earlier studies [3], [5].

2.1.4 Procedure

Following informed consent, physiological sensor attachment and completion of a pre-experimental questionnaire about safe driving (safety protocol), participants were trained in the n-back task until they met minimum performance. Each participant's baseline performance on the n-back was subsequently assessed at each of the three demand levels with 2-minute breaks between each level. Then, participants received about 20 minutes of urban road driving experience and adaptation time on the instrumented vehicle. The highway driving experiment begins when a subject is confident in safe driving with the instrumented vehicle. In a main experiment session, participants drove in good weather through 36km of highway for about 20 minutes. The driving road has speed limit of 100kph, two lanes in each way, and about 8km of uphill and downhill (3~5 percent slope). The time between 5 and 7 minutes was used as a single task driving reference (baseline). Thirty seconds later, 18 seconds of instructions introduced the task (0, 1 or 2-back). Each n-back period was 2 minutes in duration (four 30-second trials). Two-minute rest periods were provided before presenting instructions for the next task. Presentation order of the three levels of task difficulty was randomized across participants.

2.2 Ratings of Driving Workload

In order to screen participants who were highly influenced by environmental factors during performing n-back task, subjective driving workload rates were evaluated as follows. Four reviewers, who did not participate in the field study, rated the primary driving workload after watching forward view video clips. Two pairs of reviewers were seated in two different driving simulators; one is on a driver seat and the other on a passenger seat. The five 2-minute video clips, including baseline, three n-back tasks and recovery periods, were displayed on two 2.5m by 2.5m wall-mounted screens at a resolution of 1024 x 768. Each video clip was played, paused, and resumed at every 10 seconds for rating the twelve ten-second segments. For reminding the base score of driving workload, two workload anchor pictures that indicated 2 and 6 were located on the dashboard.

2.3 Dependent Variables

2.3.1 Secondary Task Performance Measures

Error rates (ER) on the n-back were used to confirm the extent to which different conditions represented periods of higher cognitive workload. The error rate is a percentage of the times when subjects answer wrong numbers or give no answer to the numbers presented to them during the n-back experiment. It can be assumed that a higher error rate indicates higher cognitive load.

2.3.2 Driving Performance Measures

Longitudinal and lateral control ability was considered as driving performance measures for indicating the difficult level of cognitive workload. In order to assess the longitudinal control performance, mean speed (SPD) and speed variability that is expressed as the standard deviation of speed (SDSPD) were selected, because some drivers have been observed performing compensatory behaviors, e.g., reducing their speed to manage the increasing workload [8]. For the lateral control ability, steering wheel reversal rate (SRR) were selected. SRR was calculated by counting the number of steering wheel reversal from the 2Hz low pass filtered steering wheel angle data per minute. Due to the factor that cognitive secondary tasks yield increased steering activity, mainly in smaller steering wheel movements, the fine reversal angles, which have more than 0.1 degree of the gap size, were counted. Although the standard deviation of lane position (SDLP) is one of the most frequently used driving performance measure, it was not used in this study due to a technical problem.

2.3.3 Physiological Measures

As shown in Table 1, six physiological measures that consist of three cardiovascular activity-based and three electrodermal activity-based measures were used. For the cardiovascular measures, mean heart rate, heart rate variation, and delta HR were considered. Mean heart rate (HR) was calculated by inverting Inter-Beat Interval (IBI) that was computed using the Librow's R-peaks detection algorithm (Librow^{TM}, Ukraine). Heart rate variation, which was calculated by standard deviation of heart rate (SDHR), was considered because variation in the inter-beat interval is a physiological phenomenon under different cognitive workload. In order to reduce individual differences, delta HR (ΔHR), which was calculated by subtracting baseline heart rate, was used. For the electrodermal measures, Skin Conductance Level (SCL) was measured with a constant current configuration and non-polarizing, low-impedance gold-plated electrodes. Sensors were placed on the underside of the outer flange of the middle fingers of the non-dominant hand without gel. Average, standard deviation, and delta SCL was calculated from the measured SCL values.

2.3.4 Eye Behavior Measures

Cognitive workload can be identified through changes in eye behaviors, for example, blink rates, pupil diameter, dwell times, characteristics of saccadic movements, and the size of the visual field. This study considered five eye behavior measures including mean and standard deviation of horizontal and vertical gaze (HG, VG, SDHG, SDVG) and blink frequency (BF). Before calculating eye-movement measures, raw gaze data were filtered with the following criteria [5]: 1) the FaceLAB's automated gaze quality index for the left and right eyes was categorized as optimal, 2) the x-axis position was between -1.5m and +1.5m, the y-axis position was between -1.0m and +1.0m, and 3) the data point was contained within a set of six valid measurements. For the eye blink frequency (BF), raw data of each period were used for calculating mean values.

2.4 Data Analysis

Statistical comparisons of the objective measures were computed using SPSS version 17. Comparisons were made using a repeated-measures general linear model (GLM) procedure. A Greenhouse-

TABLE 1 Comparison of the Sensitivity of Cognitive Workload Measures

Methods	Measures	Descriptions	Mean (S.D.)					Main Effect	Pair-wise Significance						
			Baseline	0-Back	1-Back	2-Back	Recovery		BL-2B	BL-1B	BL-0B	0B-1B	1B-2B	RC-BL	RC-0B
Secondary Task	ER(single)	Error rate of secondary task scores (%)		0.00(0.00)	0.74(1.96)	3.33(6.52)		0.111				0.164	0.184		
	ER(dual)			0.00(0.00)	1.30(2.07)	4.79(7.91)		**0.049**				**0.029**	0.097		
Driving Performance	SPD	Mean speed (kph)	98.47(6.91)	95.44(6.88)	94.41(4.83)	93.07(7.86)	98.18(7.79)	**0.012**	**0.006**	**0.009**	0.084	0.533	0.444	0.416	0.106
	SDSPD	Standard deviation of speed (kph)	6.03(1.93)	5.33(2.35)	5.95(2.11)	4.79(2.25)	5.57(2.94)	0.479	0.126	0.919	0.353	0.482	0.135	0.538	0.767
	SRR	Steering wheel reversal rate (rev. counts/min)	63.88(11.11)	62.85(11.20)	65.98(13.02)	67.76(14.33)	59.85(13.26)	**0.012**	**0.095**	0.170	0.409	0.980	0.436	0.078	0.202
	MSRR	Modified SRR with adjusted baseline	59.85(13.26)	62.85(11.20)	65.98(13.02)	67.76(14.33)	59.85(13.26)	**0.010**	**0.018**	**0.004**	0.202	0.980	0.436	0.078	0.409
Physiology	HR	Mean heart rate (beats/min)	81.33(8.47)	84.26(8.55)	85.03(8.95)	88.83(9.28)	81.32(8.61)	**0.003**	**0.004**	**0.006**	**0.001**	0.319	**0.012**	0.985	**0.001**
	SDHR	Standard deviation of heart rate (beats/min)	3.71(1.01)	4.03(1.52)	4.55(1.92)	5.25(2.28)	3.74(1.02)	**0.047**	**0.025**	0.117	0.408	0.064	0.066	0.956	0.609
	ΔHR	Heart rate difference (beats/min)	0.00(0.00)	2.93(2.19)	3.70(3.79)	7.50(7.24)	0.00(1.83)	**0.003**	**0.004**	**0.006**	**0.001**	0.319	**0.012**	0.985	**0.001**
	SCL	Mean skin conductance level (micromhos)	11.10(3.00)	11.56(3.46)	11.41(3.49)	11.75(3.60)	11.17(3.04)	0.434	0.183	0.448	0.281	0.781	0.176	0.842	0.275
	SDSCL	Standard deviation of skin conductance level	0.42(0.24)	0.06(0.48)	0.48(0.30)	0.57(0.58)	0.56(0.50)	0.634	0.300	0.502	0.156	0.448	0.611	0.373	0.631
	ΔSCL	Skin conductance level difference (micromhos)	0.00(0.00)	0.46(1.40)	0.31(1.37)	0.64(1.57)	0.07(1.12)	0.433	0.183	0.448	0.281	0.781	0.176	0.842	0.275
Eye Behavior	SDHG	Standard deviation of horizontal gaze (m)	0.47(0.16)	0.40(0.11)	0.39(0.11)	0.34(0.11)	0.45(0.10)	**0.005**	**0.004**	0.053	0.115	0.662	**0.044**	0.495	0.089
	SDVG	Standard deviation of vertical gaze (m)	0.31(0.13)	0.28(0.10)	0.27(0.10)	0.27(0.09)	0.28(0.12)	0.232	0.067	0.094	0.070	0.779	0.727	0.292	0.818
	HG	Mean horizontal gaze (m)	0.03(0.27)	-0.04(0.30)	-0.05(0.31)	-0.09(0.29)	0.01(0.31)	0.081	**0.046**	0.123	0.194	0.541	0.282	0.730	0.086
	VG	Mean vertical gaze (m)	0.58(0.39)	0.53(0.40)	0.57(0.37)	0.56(0.41)	0.52(0.34)	0.521	0.656	0.781	0.316	0.186	0.712	0.195	0.779
	BF	Blink Frequency (Hz)	0.49(0.29)	0.54(0.24)	0.58(0.26)	0.60(0.24)	0.48(0.26)	0.340	0.060	0.064	0.110	0.427	0.595	0.811	**0.041**

Geisser correction was applied for models that violated the assumption of sphericity. Post-hoc pair-wise comparisons were computed for significant effects using a least significant difference (LSD) correction.

3. RESULTS

3.1 Ratings of Driving Workload

In order to investigate the effect of the cognitive workload induced by secondary tasks, the primary driving workloads were subjectively evaluated and confirmed that the workloads across five periods, i.e., baseline, three n-backs, and recovery, were not significantly different (F(4, 37.924) =2.468, p=.078). The average driving workloads of baseline, 0-back, 1-back, 2-back and recovery were 2.6, 3.2, 3.0, 2.8 and 2.9, respectively.

3.2 Secondary Task Performance Measures

Error rates on the n-back tasks during the driving only and dual-task conditions appear in Table 1. The overall higher error rates under dual task condition mean that the demands of the primary driving task reduced the cognitive resources available to invest in the n-back. The error rates were increased as the level of cognitive task difficulty increased under both driving only conditions and the dual-task condition. However, the error rates of baseline n-back tasks were not significantly changed, because the error rates were very low across all three levels. This means all participants were highly engaged to perform the n-back tasks. For the dual task condition, the cognitive task difficulty significantly impacted on the secondary task performance.

3.3 Driving Performance Measures

As shown in Table 1, the participants significantly decreased vehicle speed as the level of cognitive task difficulty increased. The mean speed profiles showed a simple correlation with the level of cognitive workload, but the standard deviation of speed was subtle. For the lateral control ability measures, SRR measures were significantly impacted by the difficult level of cognitive workload. However, SRR in the baseline period is relatively high because the geography of baseline area was curvy downhill. Thus, Modified SRR (MSRR) was calculated by replacing the baseline with the recovery value. The MSRR profiles did show a relatively simple correlation with the cognitive level and post hoc comparisons show significant differences between some of these periods (baseline to 1-back and 0-back to 2-back). It means SRR have moderate sensitivity to differentiate the graded levels of cognitive demand and could be one of good effective cognitive measures when a driver baseline is appropriately selected.

3.4 Physiological Measures

To observe the physiological response change under different cognitive demand level, IBI, SDIBI, HR, SDHR, HRV, ΔHR, SCL, ΔSCL, SDSCL were investigated (see Table 1). Among the cardiovascular activity-based measures, IBI, HR and ΔHR were significantly impacted by cognitive task difficulty. They could differentiate most of different cognitive demand levels (baseline to 0-back, 1-back to 2-back, 0-back to recovery) except the difference between 0-back and 1-back. The main effect of cognitive demand also significantly impacted on SDHR but its pair-wise significance was limited. For the electrodermal activity-based measures, all SCL related measures were not significantly impacted by cognitive workload. These results are inconsistent with earlier findings. The reason is unclear at the moment and more careful review of experimental settings such as sensor attachment and sample differences is required.

3.5 Eye Behavior Measures

To observe the eye behavior change under different cognitive demand level, SDHG, SDVG, HG, VG, and BF were examined. Two eye behavior measures including SDHG and BF were significantly impacted by cognitive demand. SDHG could differentiate higher cognitive demand levels (baseline to 2-back, 0-back to 2-back, 1-back to 2-back), but the sensitivity of BF was limited. In this study, the sensitivity of SDHG is slightly different from Reimer's results [5] and this will be discussed in discussion section. The main effect of cognitive demand did not significantly impact on HG and VG in this study.

4. DISCUSSION

N-back error rates from both the non-driving and dual task period were increased in difficulty across the task levels. Coincident with this, several cognitive measures in multiple domains, including HR and ΔHR in physiological domain, SDHG in eye behavior domain, and SPD and MSRR in driving performance domain, showed an unambiguous increase in mean value for each level of heightened demand.

In the physiological measurement domain, patterns of change in heart rate under added cognitive demand was consistent with the earlier findings of Mehler et al. [3] except pair-wise significance between 0-back and 1-back. The difference between Mehler's results and this study could be caused by the environmental factors. As shown in Table 1, the increment in error rates between 0-back and 1-back was very small in both of non-driving and driving conditions, i.e., 0.74% under non-driving condition and 1.30% under dual-task condition. It means the difficult level of 1-back was slightly higher than that of 0-back and overall workload can be easily changed by environmental factors. In this study, the average driving workload during 0-back (workload rating: 3.2) was higher than that of 1-back (workload rating: 3.0). Thus, it can be speculated there was no difference between 0-back and 1-back period, because the combined cognitive workload was almost same due to the environmental factors. Although the driving workload induced by environmental factors was not reported in Mehler's study, the added cognitive demand between 0-back and 1-back seems to be limited to represent low and moderate cognitive demands.

In the eye behavior measurement domain, the results on horizontal eye movement were similar to the findings of Reimer et al. [5], but the patterns in gaze constriction and sensitivity in low and moderate demand were slightly different. Reimer's results showed that horizontal gaze concentration constricted in a relatively linear fashion and bottom out in 2-back. In this study, however, the highest constriction appeared in 2-back. This variation in pattern between the two studies seems to be caused by variability in the samples, i.e., the sample of 108 individuals was equally balanced by gender and across three age groups: 20 to 29 (*M*=24.6, *SD*=2.7), 40 to 49 (*M*=44.5, *SD*=3.0), and 60 to 69 (*M*=63.3, *SD*=3.1) in Reimer's study. On the other hand, the difference in sensitivity between 0-back and 1-back can be caused by the driving workload difference between the 0-back and 1-back periods as mentioned before.

In summary, this study provides general understanding of various measures for detecting the difficult levels of driver's cognitive demand. The results suggested that the patterns of change in HR, SRR, and SDHG with increasing mental workload showed near linear correlation. Among these effective cognitive measures, physiology, especially mean heart rate, showed the most sensitive response and seems to be the best independent indicator of changes in cognitive workload. Other options besides heart rate, SDHG in eye movement and SRR in driving performance measures can be used for detecting the presence of cognitive workload. Especially, SDHG could be considered one of the most useful measures in the eye behavior domain, because the vision-based approach would be capable to detect not only cognitive demand with reasonable sensitivity but also visual distraction with high accuracy. Nevertheless, the steering wheel reversal rate (SRR) is highly recommended to use for discriminate moderate level of cognitive demand, because SRR could be collected through the easiest and less expensive way. The steering reversal rate in the driving performance domain can be a commonly used measure by combining with the other domains' measures. These measures can be used for evaluating cognitive workload associated with voice interaction system, represents a potential distraction from the driving task.

5. ACKNOWLEDGMENTS

This research was supported in part by Daegu Gyeongbuk Institute of Science and Technology (DGIST) Research Program of the Ministry of Education, Science, and Technology (MEST) (Project No. 12-S-01), and Establishment Program of Industrial Original Technological Base of the Ministry of Knowledge Economy (MKE) (Project No. M0000009).

6. REFERENCES

[1] Torkkola, K., Massey, N., and Wood, C. 2004. Driver Inattention Detection through Intelligent Analysis of Readily Available Sensors. In *Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems* (Washington D.C., USA, October 03-06, 2004). ITSC2004. IEEE, New York, NY, 326-331.

[2] Son, J. and Park, S.W. 2011. Cognitive Workload Estimation through Lateral Driving Performance. In *Proceedings of the 16th Asia Pacific Automotive Engineering Conference* (Chennai, India, October 06-08, 2011). APAC16. SAEINDIA, India, SAE2011-28-0039.

[3] Mehler, B., B. Reimer, and J. F. Coughlin. 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. *Journal of the Human Factors and Ergonomics Society*, 54, 3 (June 2012), 396-412.

[4] Engström, J. A., Johansson, E., and 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 (Apr. 2005), 97-120.

[5] Reimer, B., Mehler, B., Wang, Y., and 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. *Journal of the Human Factors and Ergonomics Society*, 54, 3 (June 2012), 454-468.

[6] Victor, T. W., Harbluk, J. L. and Engström, J. A. 2005. Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transportation Research Part F: Traffic Psychology and Behaviour.* 8, 2 (Apr. 2005), 167-190.

[7] Lin, B. T. W., Green, P., Kang, T., and Lo, E. 2012. *Development and Evaluation of New Anchors for Ratings of Driving Workload*. UMTRI-2012-14. Michigan Center for Advancing Safe Transportation throughout the Lifespan.

[8] Son, J., Reimer, B., Mehler, B., Pohlmeyer, A. E., Godfrey, K. M., Orszulak, J., Long, J., Kim, M. H., Lee, Y. T., and Coughlin, J. F. 2010. Age and cross-cultural comparison of drivers' cognitive workload and performance in simulated urban driving. *International Journal of Automotive Technology*, 11, 4, (Aug. 2010), 533-539.