

Bench-Marking Drivers' Visual and Cognitive Demands: A Feasibility Study

Yu Zhang

DENSO International America, Inc.

Linda Angell

Touchstone Evaluations, Inc.

Silviu Pala

DENSO International America, Inc.

Ifushi Shimonomoto

DENSO Corporation

ABSTRACT

Objective tools that can assess the demands associated with in-vehicle human machine interfaces (HMIs) could assist automotive engineers designing safer interaction. This paper presents empirical evidence supporting one objective assessment approach, which compares the demand associated with in-vehicle tasks to the demand associated with "benchmarking" or "comparison tasks". In the presented study, there were two types of benchmarking tasks-a modified surrogate reference task (SuRT) and a delayed digit recall task (n-back task) - representing different levels of visual demand and cognitive demand respectively. Twenty-four participants performed these two types of benchmarking tasks as well as two radio tasks while driving a vehicle on a closed-loop test track. Response measures included physiological (heart rate), glance metrics, driving performance (steering entropy) and subjective workload ratings. Results suggested that multiple types and levels of benchmarking tasks can be used to give insight about reference demand levels for real, in-vehicle tasks. The four levels of visual-loading task (the modified SuRT) had graded responses in glance measures, steering entropy and subjective workload ratings. The cognitively-loading task (the n-back task) generated scalable responses in heart rate and subjective workload ratings. The range of visual and cognitive demands from the radio tasks could be usefully compared to the two benchmark tasks. Age was also found to be a significant mediating factor for secondary-task-associated workload based on all response measures. Results also suggested driving performance and physiological measures extend the understanding from benchmarking tasks beyond that yielded by subjective workload ratings.

CITATION: Zhang, Y., Angell, L., Pala, S., and Shimonomoto, I., "Bench-Marking Drivers' Visual and Cognitive Demands: A Feasibility Study," *SAE Int. J. Passeng. Cars - Mech. Syst.* 8(2):2015, doi:10.4271/2015-01-1389.

INTRODUCTION

Based on the data available in the 1990s, researchers estimated that drivers devote 30-50% of their visual attention to secondary tasks that are not directly related to driving [1] [2]. Today, drivers' attention devoted to secondary tasks may become even greater than before due to the popularity of consumer electronics. The need to provide both connectivity and safety to today's driver presents an enormous challenge to the automotive industry. One holistic solution for this challenge would be to minimize driver distraction and enhance driving safety at the design phase. Such an approach creates an immediate need for techniques that can objectively evaluate the demand associated with human machine interfaces (HMIs).

Many prior research efforts approached HMI demand evaluation through direct measurement methods by using performance-based indices, eye behavior measures, self-reported ratings, and

physiological metrics. Some utilized surrogate measurement methods, which emulate certain aspects of driving demands (e.g., the visual occlusion technique and lane change test). Others focus on drivers' situation awareness and responsiveness to external events on or near the roadway [3] [4]. Each of the three approaches provides unique perspectives on the demand and impact associated with in-vehicle HMIs and the possibility of applying absolute criteria in assessment. However, individual differences (e.g., skill, age, style) in managing HMI tasks have been found to be an important mediating factor during demand evaluation [5]. Hence, complementary assessment methods are still needed to integrate available metrics and incorporate consideration of individual differences.

The Alliance of Automobile Manufacturers provided one reference task method of assessment. In this method, performance of one type of radio tuning task while driving is used as a benchmark for an

“allowable absolute influence of a secondary task” [6]. This method consists of comparing driver performance during an HMI task relative to driver performance when tuning the radio. Based on an extension of this approach, it may be possible to use well-understood or structurally-designed reference tasks as synthesized “rulers” of HMI demand of different types. Since such “rulers” would not require predefined criteria based on absolute metrics, they would also allow integration of individual differences. In addition, research by Reimer, Mehler and their colleagues has revealed graded physiological metrics and visual behavior changes when observing drivers performing during a delayed digit recall task (n-back) presented at four different levels of difficulty during both simulated and on-road driving [7][8][9]. Their findings suggest that “benchmarking tasks” based on carefully-designed artificial tasks may be useful in gaining insight into “which” driver resources are involved in an HMI interaction, and “how much” of certain resources are involved. This approach requires the use of artificial tasks presented at multiple levels of difficulty.

The focus of this study is to assess the feasibility of using “benchmarking” tasks or “reference tasks” as “rulers” for assessing the visual and cognitive demand of in-vehicle HMI tasks. A modified surrogate reference task (SuRT) with four different levels and a delayed digit recall task (n-back task) with three variations served as references for visual demand and cognitive demand respectively (see detail in Methods section). Particularly, this study addresses four questions. **Q1.** Do the measurements gathered from the test track replicate similar measurements in previous n-back studies? **Q2.** Is it possible to assess the performance baseline of each participant by using multiple levels of reference tasks? **Q3.** Do direct driving performance measures and physiological measures provide complementary assessment of driver workload for different types of reference tasks? **Q4.** Are there any mediating factors which affect drivers' responses to the demand associated with secondary tasks, and should these be considered in a future embodiment of the “benchmarking” method?

METHOD

Participants

In this study, 24 volunteers completed all required tasks. They were recruited from a rural area near Detroit, Michigan by Morpace Inc. There were four age groups: 1) 18-24 years old, 2) 25-39 years old, 3) 40-54 years old, and 4) above 55 years. Each age group had 6 participants. Within each age group, there were an equal number of males and females, except for the 40~54 year old age group. In the 40~54 age group, there were 2 females and 4 males. All participants were active drivers, driving at least 75 miles per week. All participants were considered to have sufficient driving experience (i.e., more than 2 years). In terms of educational level, two participants were high school graduates, while most of the participants (54.17%) had received some college or post high school training; and 37.5% of the participants were college graduates or had received higher levels of education. Table 1 summarizes mean and range of age, driving experience and frequency based on four age groups and gender.

Table 1. Mean age, driving experience, driving frequency summary based on age group and gender

Group >	Gender	18~24 years old	25~39 years old	40~54 years old	Above 55
Age	Male	22.7 (22 ~ 23)	29.3 (25 ~ 36)	48.5 (46 ~ 53)	66 (56 ~ 75)
	Female	20 (18 ~ 22)	32 (31 ~ 33)	41.5 (40 ~ 43)	59 (56 ~ 65)
Experience	Male	6.7 (6 ~ 7)	13.3 (9 ~ 20)	32.5 (30 ~ 37)	50.7 (40 ~ 61)
	Female	4.3 (2 ~ 6)	15.3 (14 ~ 17)	25.5 (24 ~ 27)	43 (40 ~ 49)
Miles/week	Male	383 (150 ~ 500)	367 (100 ~ 700)	275 (50 ~ 500)	223 (120 ~ 350)
	Female	175 (75 ~ 300)	91.7 (75 ~ 100)	225 (200 ~ 250)	270 (110 ~ 400)

Benchmarking Tasks and Real In-Vehicle Tasks

Modified SuRT Task

The visual benchmarking task was based on the Surrogate Reference Task (SuRT), originally developed as part of Advanced Driver Assistance Metrics (ADAM) project [10]. This task requires a driver to look for a target circle on a screen of distractors, and select the portion of the screen on which it appears. It has a number of variations [11] [12] [13]. The version used in this study most closely resembles the version adopted by Wynn and Richardson [12]. The target circle of this task was always 72 arc minutes in diameter from the 50th percentile driver view. The target was presented among an array of distractor circles that vary in size, depending on the task difficulty level being presented. The modified SuRT task used here consisted of four test levels, corresponding to low, medium, high, and extreme levels of visual demand: 1) Easy-Level 1; 2) Moderate-Level 2; 3) Hard- Level 3; 4) Extreme- Level 4. Figure 1 below shows an example of the target circle in the array of level 1 (left side of figure) and of level 4 (right side of figure) distractor circles. Once the array of circles is presented, participants must search for and find the target circle. For this study, the SuRT task was presented on a touchscreen (a Microsoft Surface).

Participants were asked to indicate the location of the target circle by touching the half of the screen on which the target was located. (This response-method was held constant across levels of distractor size). The selected half of the screen was highlighted in light gray once the participant touched the screen. Participants were then asked to touch again (a second time) on the same half of the screen to confirm their selection (see Figure 2). Once the SuRT system received the confirmation touch, it immediately proceeded to present the next trial. Participants were given 30 seconds to complete as many target identifications as possible.

Delayed Digit Recall Task (n-back)

This task is a cognitive benchmarking task that requires a driver to remember numbers that are spoken aloud (like those in a phone number). The implementation of the n-back task follows Mehler et al.'s procedure [14]. Task difficulty was varied by changing how long the numbers must be held in memory (while new numbers are still being spoken). For this task, a string of numbers is read aloud that participants must listen to and repeat back following specific rules. At the easiest level (0-back), participants are required to

respond to each of the randomly-ordered auditory stimuli (single digits 0-9) by immediately repeating out loud the last number presented in the string they are hearing. At a moderate demand level (1-back), participants must respond with the next-to-last stimulus that was presented in the string they are hearing. And at the most difficult level (2-back), participants must respond with the second-to-last stimulus in the string they are hearing. The n-back task used in this study was presented at all three levels of difficulty, corresponding to low (0-back), medium (1-back), and high levels (2-back) of cognitive demand. In each experimental trial, participants responded to 2 series of 10-digit numbers. The total task duration of each trail was 1 minute.

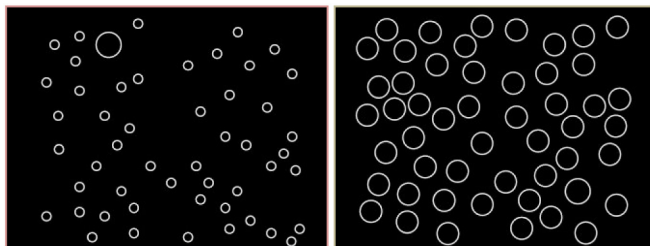


Figure 1. SuRT task examples: Left image shows target circle 1.40 cm in diameter among an array of distractor circles 0.50 cm in diameter - Level 1 difficulty. Right image shows target circle 1.40 cm in diameter among an array of distractor circles 1.25 cm in diameter - Level 6 difficulty.

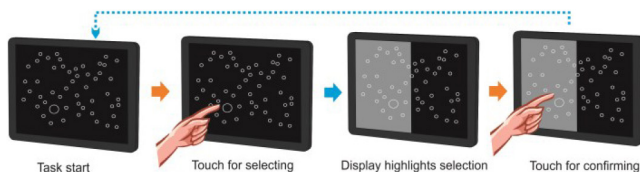


Figure 2. Illustration of SuRT task action sequence.

Radio-Tuning Task

There were two levels of radio task. At an easy level, participants were asked to tune the radio to a specific FM station starting from another FM station that was playing. At a hard level, participants were asked to select the radio function from the home page (an action intended to correspond with “turning the radio on” in the Alliance reference task), select a band and then tune to a specific station. (The easy level of radio task was adopted from the Driver Workload Metrics Projects by CAMP [5]. The hard level of radio task was adapted from the Alliance recommended procedure for the radio tuning reference task [6].)

Test Track & Apparatus

The test track was located at Fowlerville, Michigan. It was a 3-mile oval track with two, 12-ft. wide lanes. The road surface was asphalt with no banking in straight segments. The south curve was 1,000 ft. in radius, and the north curve had a decreasing radius from 1,500 ft. to 1,000 ft. The straightaway length was 4,500 ft.

A full-size sport utility vehicle (Model Year 2013) was used for testing purposes on the oval track. As shown in Figure 3, a Microsoft Surface (1st generation) had been securely mounted on a rigid mount that allowed presentation of modified SuRT tasks. The mount was removed when the driver performed radio tasks. The radio system in car was used as the platform for all radio tasks.

The main data collection unit was mounted behind the second row of seats. This unit consisted of a main data acquisition computer, a SmartEye eye tracking system, a Zynex medical-grade physiological data recorder, a 5-camera observation system, an omnidirectional microphone and a CAN message reader. It recorded glance measures as well as physiological measures (e.g., heart rate and skin conductance). It was also used to record driving performance measures acquired from the vehicle, such as steering angle, vehicle speed and interactions with HMI devices.



Figure 3. Microsoft surface for SuRT with mounts in the instrumented vehicle

Experiment Design

This study followed a $4 \times 2 \times 9 (\times 2)$ repeated-measures factorial design with two between-subject factors (age group and gender) and one within-subject factor (secondary task). That is, each participant performed 9 secondary task conditions, including: 2 levels of radio task, 4 levels of modified SuRT task and 3 levels of n-back task. Each secondary task condition was performed twice. To avoid fatigue, each major type of secondary task had its own testing block. Participants had long breaks between test blocks, and their levels of comfort and fatigue were regularly monitored throughout the experimental procedure. In addition, a “just driving” condition was implemented to collect performance baselines for each major type of secondary task.

Response Measures

Secondary Task Performance

Task performance measures were determined based on characteristics of secondary tasks. For the SuRT task, response accuracy and average response times were recorded. For the n-back task (forced-pacing), response accuracy was collected. For radio tasks, task completion times were recorded.

Subjective Workload Rating

Participants reported their subjective perception of workload for each secondary task by using a global workload rating method [15]. This method required participants to rate their perceived workload using a 0 (low) to 10 (high) scale for the task they just completed. (The scale allowed half point resolution.)