U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

NHTSA
www.nhtsa.gov

DOT HS 810 635

November 2006

Driver Workload Metrics Project

Task 2 Final Report

CAMP
Driver Workload Metrics

IVI Light Vehicle Enabling Research Program

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT HS 810 635		

4. Title and Subtitle	5. Report Date
Driver Workload Metrics Task 2 Final Report	November 2006
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
L. Angell, J. Auflick, P.A. Austria, D. Kochhar, L. Tijerina, W. Biever, T. Diptiman, J. Hogsett, and S. Kiger	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Crash Avoidance Metrics Partnership (CAMP) 39255 Country Club Drive, Suite B-30 Farmington Hills, MI 48331	11. Contract or Grant No. DTFH61-01-X-00014

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
National Highway Traffic Safety Administration U.S. Department of Transportation 400 Seventh Street SW. Washington, DC 20590	Final Research Report April 2001 – June 2005
	14. Sponsoring Agency Code

15. Supplementary Notes

The Crash Avoidance Metrics Partnership (CAMP) is a partnership established by Ford and GM to undertake joint pre-competitive work in advanced collision avoidance systems. CAMP was established to accelerate the implementation of automotive crash avoidance countermeasures and define pre-competitive enabling elements.

16. Abstract

This report presents the results of a study of driver workload associated with use of in-vehicle systems while driving. The purpose of the study was to develop performance metrics and test procedures to assess the visual, manual, and cognitive aspects of driver workload. A second objective was to develop a "toolkit" of evaluation methods to enable developers to manage the driver workload implications of future products. Driver performance data was collected in three venues: in the laboratory (N = 57 drivers), on an interstate highway (N = 108 drivers), and on a test track (N = 69 drivers). Twenty-two in-vehicle tasks, plus the task of just driving, were examined under a variety of experimental conditions. Analysis of the data focused on selecting metrics that were repeatable, had predictive validity, and discriminated higher workload from lower workload tasks. The results indicated that task-induced workload on driving performance is multidimensional in nature. The effects observed depended on the characteristics of the task. Visual-manual tasks exhibited fundamentally different performance profiles than auditory-vocal tasks. The differences included both the dimensions affected and the magnitude of effects. Driving performance metrics identified for the workload evaluation toolkit included metrics from lateral and longitudinal control, object and event detection, and eyeglance behavior. Laboratory tools included an activity time model, a static task time measurement tool, a visual occlusion method, a peripheral detection task used with and without a part-task driving simulator, and a Sternberg memory task assessment tool.

17. Key Words	18. Distribution Statement
human factors, driver workload metrics, driver distraction, in-vehicle tasks, visual demand, cognitive load, workload assessment methods, vehicle control metrics, driver eyeglance behavior	No restriction. This document is available to the public from the sponsoring agency at the Web site http://www.cflhd.gov.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	1006	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Table of Contents

List of Figures

List of Tables

Abstract

This report presents the results of a study of driver workload associated with use of in-vehicle systems while driving. The purpose of the study was to develop performance metrics and test procedures to assess the visual, manual, and cognitive aspects of driver workload. A second objective was to develop a "toolkit" of evaluation methods to enable developers to manage the driver workload implications of future products during all stages of the design process.

In the study, driver performance data were collected in three venues: in the laboratory, on an interstate highway, and on a test track. Two hundred thirty-four licensed drivers were recruited for participation in the study. Each driver participated in only one testing venue. The participants ranged in age from 21 to 79 and were balanced by gender. In each venue, the participants performed in-vehicle tasks under a variety of experimental conditions. Twenty-two in-vehicle tasks were examined in this study. In addition, a two-minute segment of just driving was performed under the same conditions for comparison purposes.

The laboratory phase investigated the use of computer-based methods for assessing driver workload in lieu of actual driving. Methods examined in the laboratory included a peripheral detection task used with and without a fixed-based driving simulator, a visual occlusion technique, a memory task, static task completion technique, and subjective assessments of overall workload, situational awareness and multitasking difficulty.

In the on-road and test track phases, test participants drove an instrumented car while performing the in-vehicle tasks. The car driven by the participant was the center car of a three-vehicle platoon operated as a single testing unit. The leading and following cars in the platoon were used to provide a car-following scenario as well as event detection scenarios that featured lead-vehicle deceleration, lead-vehicle center high-mounted stoplight activation, and follow vehicle turn signal activation. On-board instrumentation recorded vehicle control data, eyeglance patterns, and responses to the event detection scenarios.

In addition to driver performance testing, several analytic models were examined in the study. These included a count of the task's steps, time estimates for physical and cognitive activities in the task, and several indicators based on Multiple Resource Theory of a task's potential to compete for a driver's resources while driving.

Analysis of the driving performance data focused on the repeatability and discriminability of potential metrics from lateral and longitudinal control, object and event detection, and eyeglance data. The results indicated that task induced workload on driving performance is multi-dimensional in nature. No single metric presents a complete picture of the task effects observed in the study. Furthermore, the effects observed depend on the characteristics of the task. Visual-manual tasks exhibited fundamentally different performance profiles from auditory-vocal tasks. Differences included both the dimensions affected and the magnitude of the effects. Repeatable driving performance metrics that discriminated multitasking from just driving and hypothesized high workload from low were identified for visual-manual and auditory-vocal tasks separately.

Analysis of laboratory data featured assessments of repeatability and discriminability. In addition, laboratory metrics also were required to be predictive of driving performance data. Metrics meeting these requirements were identified for visual-manual and auditory-vocal tasks.

A toolkit of assessment techniques also was recommended. The toolkit included an activity time model as well as the laboratory methods involving static task completion, visual occlusion, peripheral detection task, memory task, and vehicle control metrics from the simulator. Instrumented vehicle metrics also were included in the toolkit for use in the latter stages of product development when pre-production prototype devices are available. Recommendations for future research are included in the report.

Executive Summary

Scope

Drivers of today's vehicles are faced with an increasing array of competing demands for their attention. The primary demand, of course, is the need to be aware of and respond to constantly changing road and traffic conditions. Secondary or competing demands are many and include in-vehicle information and communication systems. These systems range from those commonly found in vehicles to newer telematic devices such as wireless navigation systems.

The competing demands often require the driver to multitask. That is, the driver typically continues to drive while using or responding to an in-vehicle system. Multitasking increases the workload demand on drivers. If the workload demands exceed a driver's capacity in some way, then the driver's ability to drive will be degraded. If the degradation is significant, and if other contributing factors such as sudden changes in traffic or unexpected roadway objects occur, then the likelihood of a crash or near-miss increases. This logic provided the foundation for the Driver Workload Metrics project and is shown in Figure 1.

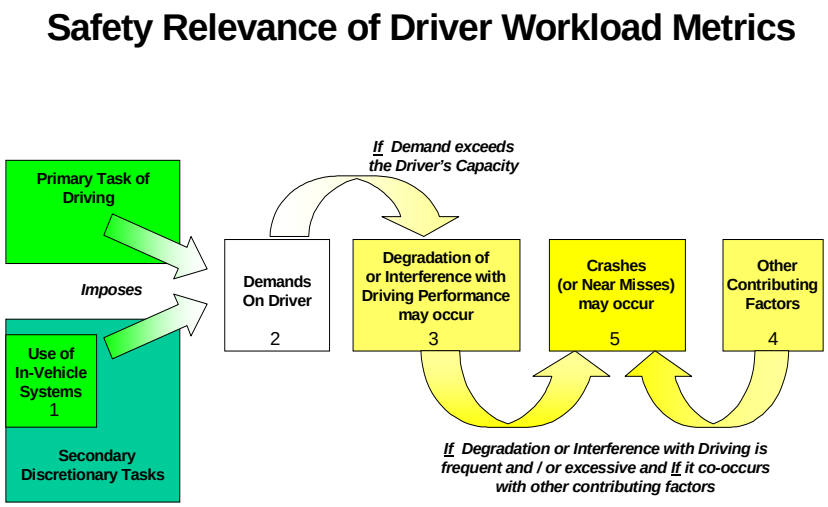


Figure 1. Safety Relevance of Driver Workload Metrics

The focus of the Driver Workload Metrics project was the linkage between the demand placed on the driver by secondary discretionary tasks (block 1) and the potential interference with driving performance (block 3) depicted in the left half of Figure 1. The project established a candidate set of performance metrics for use in evaluating the extent of any interference with driving performance resulting from secondary discretionary tasks. Of equal importance, but outside the scope of this study, is the relationship depicted in Block 4 addressing frequency of use and the conditions under which drivers engage in discretionary tasks, as well as the potential for co-occurrence (frequency, nature, and timing) of other contributing factors. A better understanding of this relationship might be obtained from naturalistic driving studies. Such studies would assist in understanding the overall relationship between secondary device usage and crash risk.

Background

The Crash Avoidance Metrics Partnership (CAMP) was formed in 1995 to facilitate cooperative pre-competitive industry/government research designed to accelerate the implementation of crash

avoidance countermeasures to improve traffic safety by defining and developing necessary pre-competitive enabling elements of future systems. The CAMP Driver Workload Metrics (DWM) Project brought together Ford Motor Company, General Motors Corporation, Nissan Technical Center North America, Inc., and Toyota Technical Center USA with the U.S. Department of Transportation to develop performance metrics and test procedures to evaluate the visual and cognitive aspects of driver workload from telematics systems. Launched in April 2001 and completed in March 2005, the research investigated both driving performance measures of driver workload taken under test track or on-road driving conditions as well as surrogates, which include models, simulations, or laboratory procedures.

Objectives

The project addressed the measurement of driver distraction related to the use of in-vehicle systems and secondary tasks. The primary goals of the project were to:

- Develop performance metrics and test procedures to assess how the workload associated with using an in-vehicle system might degrade or interfere with driving performance.

- Establish a toolkit of correlated analysis, development, and validation procedures that enable developers of telematics devices to efficiently manage the driver workload implications of future systems during all stages of the design process.

Experimental Approach

Driver performance data was collected in each of three venues using a phased testing approach: in the laboratory, on public highways, and on a test track. Based on analysis of distraction-related crash data, the driving condition selected for testing was a highway speed car following scenario on a straight level road under clear, dry, daytime conditions. Two hundred thirty-four licensed drivers were recruited for participation in the study. Each driver participated in only one of the testing venues. The sample of participants was approximately balanced by gender and age. The participants ranged from 21 to 79 years of age.

Testing took place over two consecutive days. All participants were trained in the proper method for performing each in-vehicle task prior to the start of the testing. In each venue, the participants performed the in-vehicle tasks under a variety of conditions. In addition, participants also performed a two-minute segment of just driving under the same conditions. The segment of just driving was included for comparison purposes.

The purpose of the laboratory phase of the study was to investigate the use of simulator-based, computer-based, and model-based methods that could be used for assessing driver workload in lieu of actual driving. Metrics collected in lieu of driving performance measures are referred to as “surrogates” in this study. In the laboratory phase, the test participants were asked to perform the in-vehicle tasks while seated in a driving-like environment. The testing methods examined included a peripheral detection task, a memory task, a visual occlusion technique, a peripheral detection task used in conjunction with a fixed-base driving simulator, subjective assessments of workload and multitasking difficulty, and a static task completion technique. Surrogates were subsequently correlated with driving performance metrics during the analysis phases of the project.

As part of the laboratory efforts, the project staff also developed several analytic surrogates. The analytic surrogates were outputs from models derived from a task analysis of the in-vehicle tasks used in the study. Analytic models are important tools for ergonomic analysis of product designs,

especially early in the development cycle when design changes can be made more easily than at points closer to product release. Analytic surrogates developed for each task included a count of the steps needed to complete a task, time estimates of physical and cognitive activities in the task, and several indicators of a task's potential to compete for a driver's resources (physical, cognitive, and working memory) when performed while driving.

In the on-road phase of the study, the test participants drove an instrumented car on an interstate highway. Each participant drove the center vehicle in a three-car platoon operated together as a single testing unit. This is depicted in Figure 2. During testing, the vehicles were operated at nominally 55 miles per hour with approximately 120 feet of spacing between the vehicles. Members of the project team drove the leading and following cars in the platoon. The leading and following vehicles provided the capability to present three types of object-and-event detection scenarios during testing. These scenarios were: a leading-vehicle deceleration, leading-vehicle center high-mounted stoplight (CHMSL) activation, and the following-vehicle turn signal activation. The methods used in the test track phase of the project were similar to those used in the on-road phase. The only difference was that the testing was conducted on a five-mile oval test track to allow the more demanding tasks to be evaluated in a controlled traffic environment.

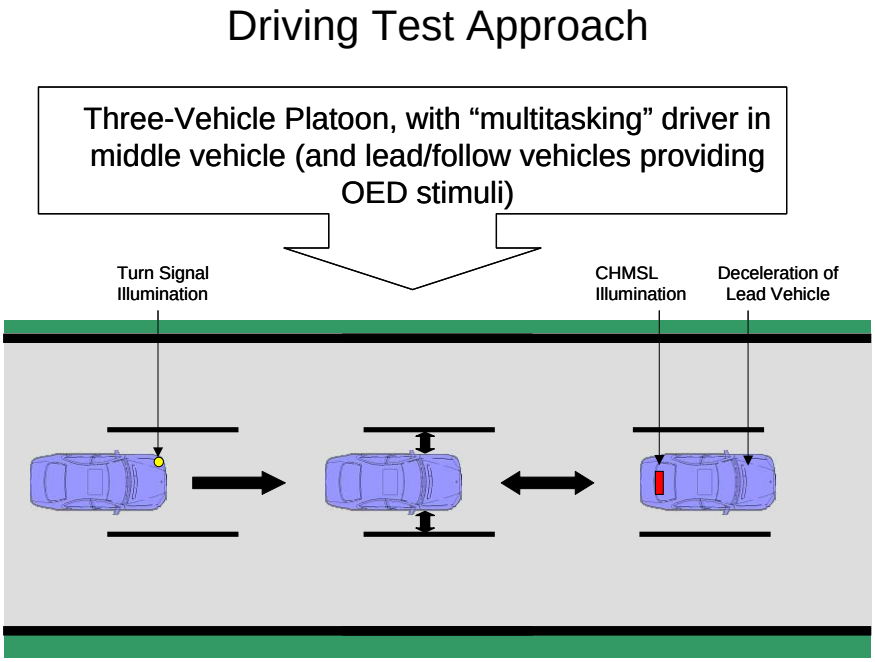


Figure 2. Driving Scenario

Task Selection

The experimental work required a set of in-vehicle tasks be selected for study. Because workload is a multidimensional phenomenon, it can affect the driver in many ways. The tasks selected needed to impose demands on the driver's input modalities (auditory or visual), output modalities (manual or vocal), and working memory (verbal or spatial) in an organized way so that their effects on driving could be examined. In addition, the tasks needed to represent device and interface types either in use today or expected in future telematics systems. Finally, the tasks needed to provide levels of expected workload difficulty that would range from low to high.

A set of tasks was selected that included conventional tasks commonly performed in vehicles today such as radio tuning, heating/air conditioning adjustment, and listening to a sports news broadcast; telematics-like tasks such as map reading, text reading, and navigation system destination entry; as well as use of a cellular telephone and tasks that were artificial to the driving environment but necessary to make sure that needed combinations of demands on driver's input and output modalities and working memory were represented. Tasks such as remembering route instructions and performing trip computations are examples of artificial tasks included in the set of tasks selected. Figure 3 presents the list of the 22 in-vehicle tasks used in the study plus the task of just driving alone. The tasks have been classified by the input and output modalities needed to perform the task: either visual input and manual output, or auditory input and vocal output. The task of just driving was included in the set for comparison purposes. The just drive task was placed in the lower workload auditory-vocal task category because its workload effects were expected to be of a cognitive nature and its duration was similar to the duration of most tasks in this category.

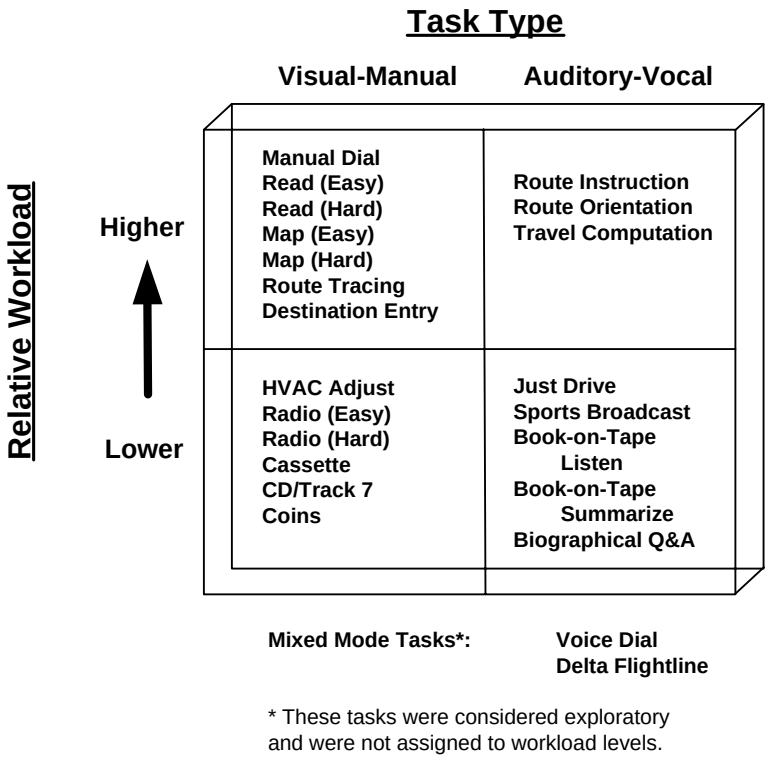


Figure 3. Tasks Studied in Project

Driving Performance Metrics

Based on a literature review of metrics with which to characterize driver workload, four safety-relevant categories of driving performance metrics were identified as important. These were:

- **Driver Eye Glancing Patterns** – visual indicators such as the number of glances, duration of glances, and the location of glances made while performing a task
- **Lateral Vehicle Control** – such as lane positioning and the number of lane line crosses

- **Longitudinal Vehicle Control** – such as speed maintenance
- **Object-and-Event Detection** – such as the percentage of missed events and response times

The instruments installed in the test vehicles were developed to continuously measure these categories of metrics while the test participant was driving.

Surrogates

Surrogates were selected for study based on several criteria. The criteria included considerations such as the aspect of human performance the surrogate would assess (like visual demand, cognitive load, or overall workload), ease of implementation, relevance to the driving task, and the need to include subjective as well as objective methodologies. The following surrogates were examined in the study.

- **Static Task Time**
  A participant performed a task without any other concurrent task or interruption. This surrogate was applied to only visual-manual tasks of variable duration. Task completion time was the metric generated by this surrogate.
- **Visual Occlusion**
  A participant performed a task wearing special goggles equipped with a computer-controlled shutter. The shutter was opened and closed repetitively throughout the task on a 1.5-second open and 2.0-second closed cycle. This surrogate was applied only to visual-manual tasks. The total shutter open time was the key metric generated by this surrogate.
- **Sternberg Memory Task**
  This surrogate involved road sign memorization and recall. While a participant performed a task, a road sign was briefly presented on a display. The participant was asked to press one pushbutton if the displayed sign was from a set of signs memorized prior to the start of the task, or a second pushbutton if not. One version of this surrogate used route junction signs while another version used route number signs. The two versions enabled the investigation of task effects on spatial and verbal working memory. This surrogate provided percent errors and response time metrics.
- **Peripheral Detection Task**
  A high-intensity spot of light was briefly projected on a screen in front of the participant during task performance. The participant activated a pushbutton in response to the light. The percent of missed events and response times were obtained from this surrogate.
- **Peripheral Detection Task with Fixed-Based Driving Simulator**
  The peripheral detection task (PDT) described above was used in conjunction with a fixed-base, part-task driving simulator from Systems Technology, Inc. The simulation involved a car-following scenario. In addition to the PDT metrics, this surrogate also produced lanekeeping and speed maintenance metrics.
- **Operator Workload Assessment**
  This subjective methodology involved a test participant rating a task on overall workload using a 100-point scale.
- **Multitasking Difficulty Assessment**
  A test participant was asked to rate how hard it was to perform a task while

driving. Ratings were obtained relative to the standard task of tuning a radio. The radio task was assigned a fixed rating value of 100. Tasks rated twice as difficult as radio tuning received a value of 200, while tasks half as difficult as the tuning task received a rating of 50, etc.

- **Situational Awareness Assessments**
  This subjective methodology involved a participant rating a task on how aware the participant felt about roadway, traffic, and other events while performing the task. The rating method used in this surrogate was the same as that used for multitasking difficulty.

Analysis

It was hypothesized that both the visual-manual and auditory-vocal tasks would interfere with driving, but in different ways. It was also hypothesized that the laboratory surrogates would exhibit similar effects from the in-vehicle tasks, and that the effects could be used to predict driving performance. The objective of the analysis phase was to identify which of the metrics were affected by the workload from the in-vehicle tasks. Because there were a large number of potential metrics that could be used to assess workload, an analysis strategy was used to screen metrics and identify the most important. To be useful in future applications, the metrics selected had to be:

- **Repeatable** – produces consistent results whenever measured as determined by split-group analysis
- **Discriminating** – distinguishes high- from low-workload tasks, or multitasking from just driving, as determined by paired comparison analysis
- **Predictive** – correlates metrics with driving performance metrics for laboratory surrogates

Results

After assessing the repeatability of candidate driving performance metrics, the remaining measures were subsequently assessed for their ability to discriminate levels of workload. Two levels of discriminability assessments were made. The first assessment compared multitasking, or performing any of the in-vehicle tasks while driving, with just driving alone. In the second discriminability assessment, the high-workload tasks were compared with the low-workload tasks. In both of the discriminability analyses, the auditory-vocal tasks were examined separately from the visual-manual tasks.

Metrics that were repeatable and that discriminated multitasking from the just drive task are shown in Table 1. Note that in the table, some metrics indicate more workload as the metric increases, while other metrics indicate more workload as the metric decreases.

Table 1. Driving Performance Metrics That Discriminate Multitasking From Just Drive

Auditory-Vocal Tasks	Visual-Manual Tasks
Duration of Glances to Road	Percent of CHMSL Events Missed
Proportion of Task Gazing at Road	Percent of Lead Vehicle Deceleration Events Missed
Percent of Follow Vehicle Turn Signal Events Missed	Percent of Follow Vehicle Turn Signal Events Missed
Number of Glances to Mirrors	All Repeatable Eyeglance Measures
Duration of Glances to Mirrors	Speed Difference
Glance Rate to Road	

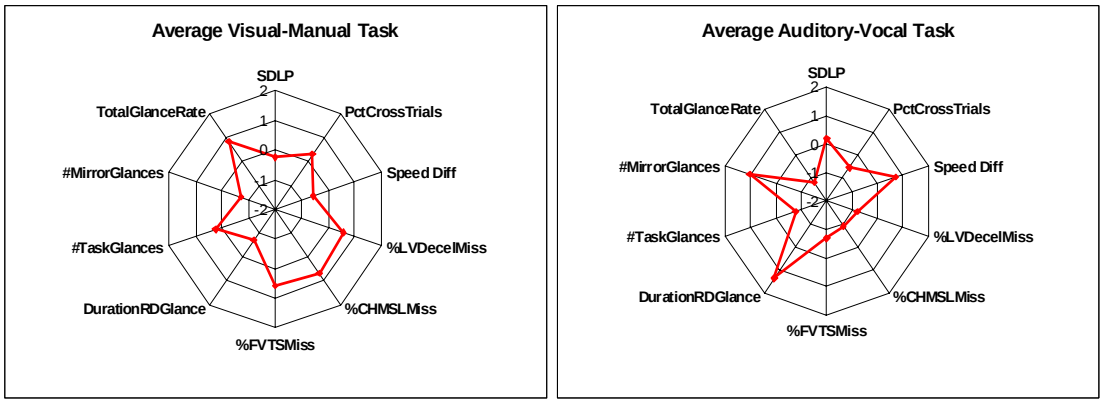
The second level of discriminability assessment compared hypothesized high-workload tasks with hypothesized low-workload tasks. The results of this assessment are presented in Table 2. In the table, no driving performance metrics were able to discriminate high- from low-workload tasks for any of the auditory-vocal tasks. However, task duration, standard deviation of lane position, speed difference, and selected eyeglance metrics were able to perform this discrimination for visual-manual tasks. The eyeglance metrics that discriminated high- versus low-workload visual-manual tasks included predominately the number of glances and duration of glances to task-related areas in the vehicle.

Table 2. Driving Performance Metrics That Discriminate Hypothesized High Workload From Low

Auditory-Vocal Tasks	Visual-Manual Tasks
None	Task Duration
	Standard Deviation of Lane Position
	Speed Difference
	Selected Repeatable Eyeglance Measures

In addition, visual-manual tasks were found to exhibit a fundamentally different performance profile from auditory-vocal tasks. The results of these findings are best represented in the multi-dimensional graphical comparison shown in Figure 4. In this figure, 10 dimensions of workload are depicted on a single graph for the average visual-manual and average auditory-vocal task. The term average task means the average across all tasks in a task type, and was computed for each metric depicted. The data presented in Figure 4 are statistically normalized values (z-scores) for each metric. Using normalized data permits a comparison of metrics with different measurement units on an equivalent basis.

The left graph in Figure 4 indicates that visual-manual tasks most affect the glancing rate, number of glances and the percent of CHMSL, follow vehicle turn signal, and lead vehicle deceleration events missed during the task. This is consistent with the driver's need to remove the eyes from the road ahead and look inside the vehicle to perform this type of task. In contrast, the auditory-vocal tasks do not require the driver to look inside the vehicle to perform the task. Consequently, the right half of Figure 4 shows much less impact on the dimensions that were affected by the visual-manual tasks. In this case, the patterns of performance during auditory-vocal tasks are observed along the dimensions represented by the number of mirror glances, duration of on-road glances, and speed difference.



Legend			
Label	**Description**	**Label**	**Description**
SDLP	Standard deviation of lane position	%FVTSMiss	Percent of follow vehicle turn signal events missed
PctCrossTrials	Percent of trials with a cross of lane line	DurationRDglance	Duration of on-road glances
SpeedDiff	Speed difference	#TaskGlances	Number of glances during the task
%LVDecelMiss	Percent of lead vehicle deceleration events missed	#MirrorGlances	Number of glances to mirrors
%CHMSLMiss	Percent of follow vehicle turn signal events missed	TotalGlanceRate	Glance rate

Figure 4. Comparison of an Average Visual-Manual Task With an Average Auditory-Vocal Task

Another important finding was that visual-manual tasks had a more pronounced effect on driving performance than the auditory-vocal tasks. This is illustrated in Figure 5 for the metrics:

- percent of task spent looking at the road;
- percent of task spent looking at mirrors;
- percent of CHMSL events detected; and
- percent of follow vehicle turn signal events detected.

For each metric in Figure 5, the just drive task is plotted along with the average auditory-vocal task and the average visual-manual task. For the percent of task spent looking at the road (upper-left graph), the auditory-vocal tasks involved approximately 7 percent more time on-road compared with just drive, while the visual-manual tasks were associated with about 40 percent less time on the road. In this case, the magnitude of the visual-manual task effect (40%) is over five times larger than the auditory-vocal task effect (7%). More subtle differences are depicted for the percent of task looking at mirrors metric (upper-right graph). This graph shows that about 14 percent of the just drive task duration was spent looking at the mirrors versus 11 percent for the auditory-vocal tasks, a difference of only 3 percent. By comparison, about 8 percent of the visual-manual task duration was spent looking at the mirrors. In this latter case, the change compared to just drive is 7 percent. The percent missed CHMSL and follow vehicle turn signal events show similar results in that the magnitude of the auditory-vocal effects are smaller than the visual-manual task effects.

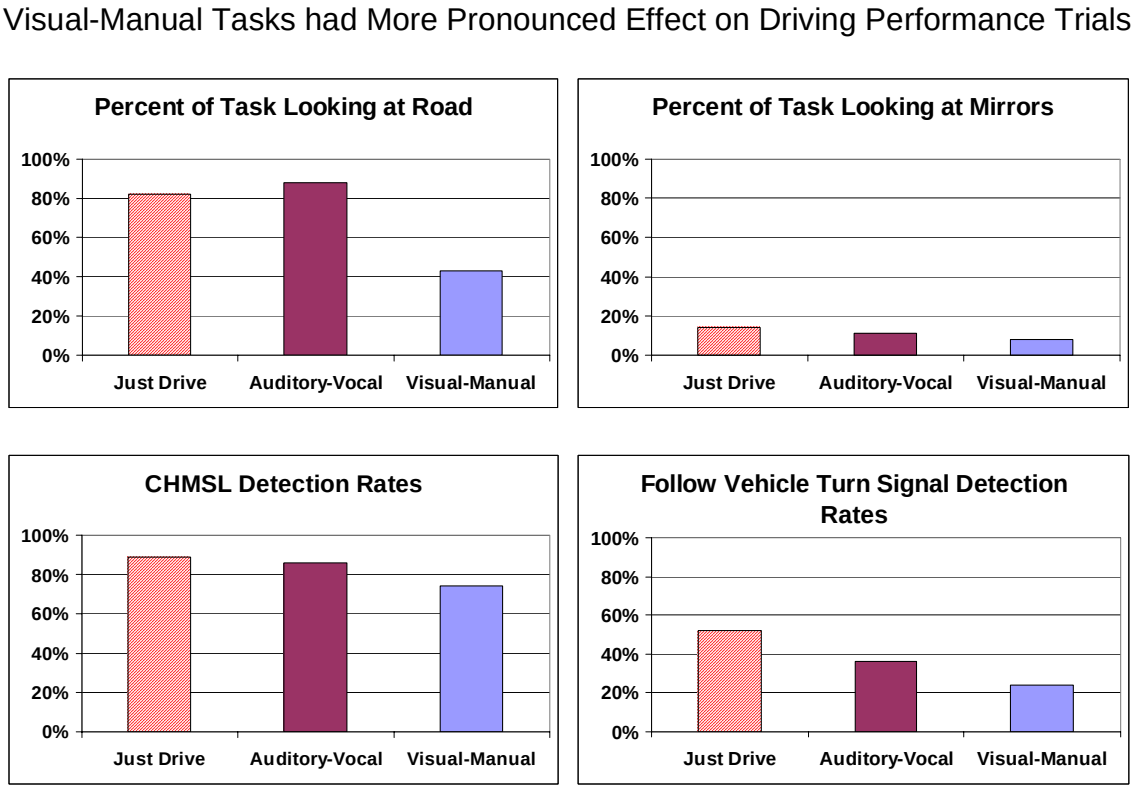


Figure 5. Illustration of the Magnitude of Effects on Driving Performance

When drivers do detect events, their glance patterns change. An example of this effect is presented in Figure 6 in which the rate of glancing to the mirrors increases following the detection of a CHMSL event. This figure presents a time series plot of multiple metrics for a portion of a task. Detection of the CHMSL event (depicted by the dark-blue line) occurred at a point near the middle of the graph. Four mirror glances (shown by the brown line) were made in the 60 seconds preceding detection of the event compared with 10 mirror glances in the 50 seconds that followed event detection. Overall, glance patterns observed in the study showed changes in frequency, rate, and duration for trials in which an event was detected versus trials in which no detection took place. The way in which glance patterns changed depended on the type of event detected. For lead vehicle deceleration events, duration of glances to the road increased and the rate of glances to task-related areas and mirrors decreased. For CHMSL and follow vehicle turn signal events, durations of glances decreased for all locations and the glance rate to the road and mirrors increased. To illustrate this point, data for mean glance duration during the follow vehicle turn signal events is presented by task in Figure 7. Shown in Figure 7 are the mean durations with and without detection of the event. This figure clearly shows the dramatic decrease in glance durations for the auditory-vocal tasks and just drive in trials with event detection. The visual-manual tasks exhibited little change in mean glance duration between trials with and without a turn signal detection.

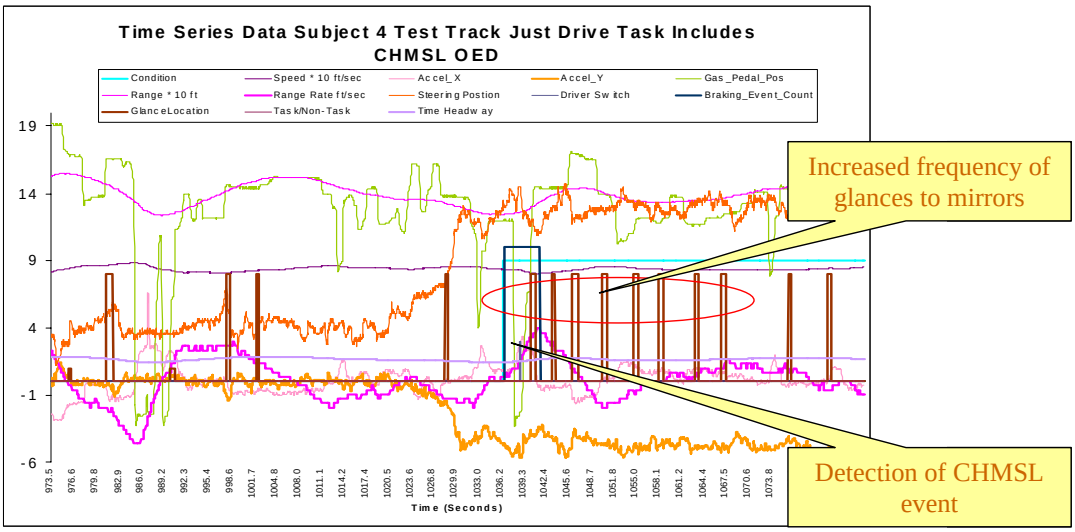


Figure 6. Example of Changes in Glance Patterns Before and After Detection of an Event

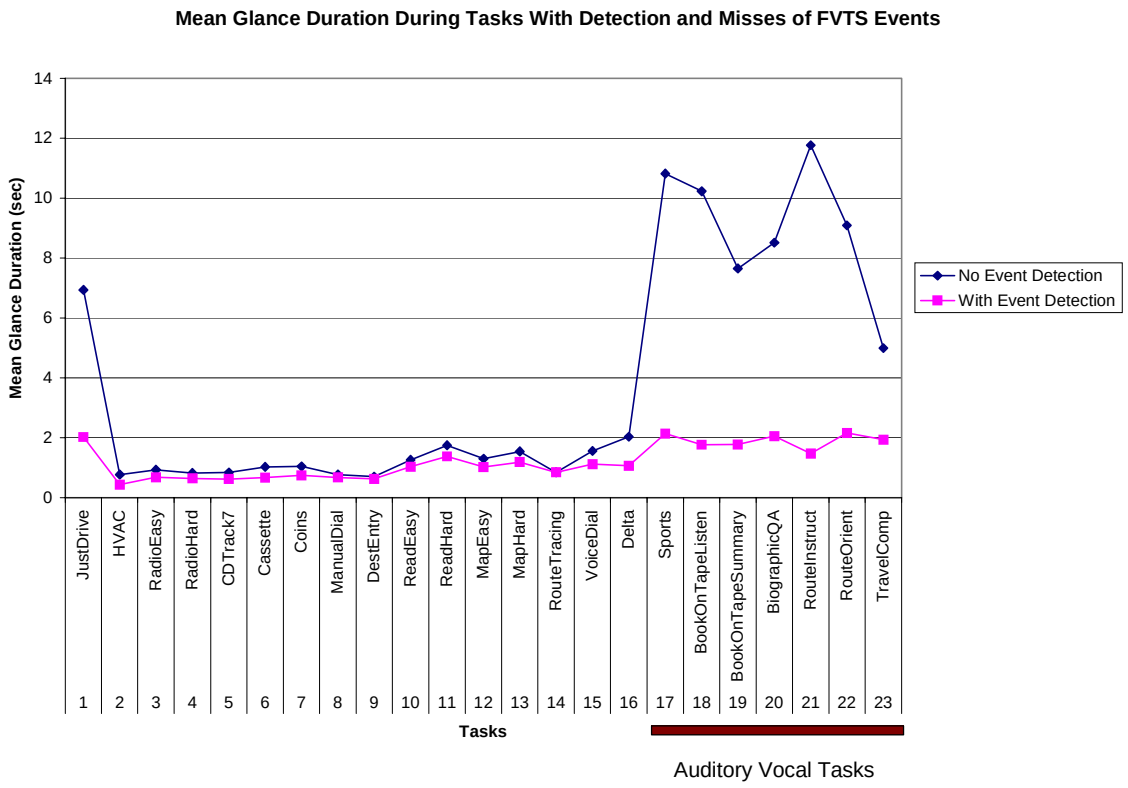


Figure 7. Comparison of Glance Duration During Tasks With Follow Vehicle Turn Signal Events

Eyeglance data examined also indicated that auditory-vocal tasks and just drive are associated with a hypothesized "shedding" of mental workload when events are detected. Theoretically, the implications of these findings are that event-detection may serve as an "attentional interrupt" for auditory-vocal tasks and the task of just driving. Following the interrupt, the driver engages in more active scanning of the road and mirrors for situational awareness. This finding is illustrated in Figure 7 by the decrease in mean glance duration for trials with event detection compared with trials involving no event detection. As glance durations decrease, glancing frequency increases. However, for visual-manual tasks, this effect was not as pronounced and appeared to occur only for the lead vehicle deceleration.

These findings have important implications for eyeglance measurement practice. When assessing the visual demands of a task, it is essential that the trials used to generate the eyeglance data for this evaluation not include the presentation of events to be detected. Including eyeglance data from event detection trials can significantly distort the assessment of the visual demand of a task. Other trials, however, should be included in the product evaluation methodology in which event detection is presented so that tasks effects on object-and-event detection are addressed. The key point is that data gathered for one aspect of task assessment should not be used for other aspects without careful consideration.

Surrogate data were also assessed in the study for repeatability, discriminability, and predictiveness. Table 3 presents the recommended list of surrogates that are repeatable, discriminate workload levels and predict (are correlated with) selected driving metrics.

Table 3. Recommended Surrogates to Predict Driving Performance Metrics

	Driving, Object-and-Event Detection, and Selected Eyeglance Behaviors						
DWM Surrogate Measure	Task Time	Driving SDLP	Driving Percent Lanex (Cross)	Driving Speed Difference	Object & Event Detection	Task Related Eye-glance Counts	Task Related Eyes Off Road Ahead
Multitasking Difficulty Scale			**				
Static Time	*					*	*
TSOT	**			*		**	**
Median STISIM Duration	**			**		*	*
Median STISIM SDLP		**	*				
Median STISIM SpeedDiff				**			
PDT-STISIM (PDTS) Percent Miss					*		
Sternberg Percent Missed Detects					**		
Sternberg Percent All Errors					**		
Sternberg Median All RT					*		
Sternberg Combined Decrement					**		

Note: * is “good” and ** is “better” in a relative sense.

Some effects were observed in the laboratory, but not on the road. For example, no object-and-event detection metrics discriminated (hypothesized) high and low workload for auditory–vocal tasks on the road, while some laboratory surrogates did. Possible explanations for this are that drivers may perceive risks differently in the lab than in real vehicles, or that the on road experiment was somehow insensitive to these effects. Until this discrepancy is better understood, judgments on task effects should not be based solely on laboratory results. Nonetheless, surrogates can be used iteratively through product development to manage workload implications of new system designs. A toolkit was defined to support this process.

Workload Evaluation Toolkit

Figure 8 presents the product development process and illustrates how evaluation tools can be utilized at different points in the cycle. Early in the development cycle, when no actual device exists (pre-prototype), workload assessments using analytic methods would be valuable. Estimates of activity completion time derived from task analysis-based models such as those developed in this study are recommended for this phase of development. Bench-testing methods would become important when initial interactive prototype devices become available. In this phase of development, recommended tools include methodologies for the static task time, visual occlusion, Sternberg memory task, and the peripheral detection task. The peripheral detection task methodology could be used either alone or with a part-task driving simulator. In the latter case, driving performance metrics would also be available to support device evaluations. Finally, Figure 8 illustrates that instrumented vehicle methods are recommended for the latter stages of development when pre-production prototypes are available. In this phase, eyeglance metrics are highly recommended among the tools used to assess final products.

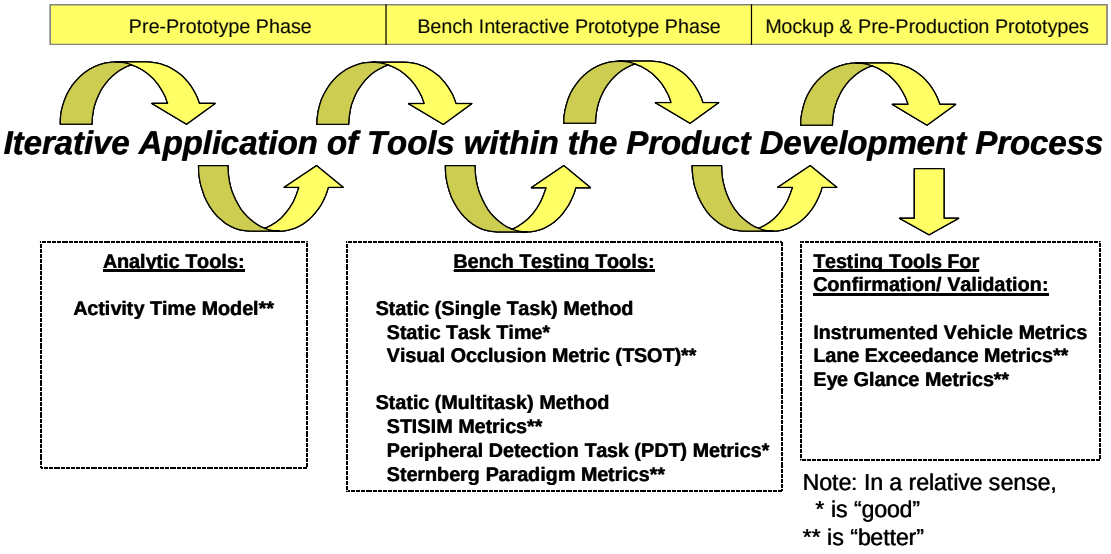


Figure 8. Driver Workload Metrics Toolkit

Conclusions

The CAMP DWM project yielded several key insights about the effects of multitasking on driving performance, including:

- Task induced workload on driving performance is multidimensional in nature. No single metric presents a complete picture of the task effects observed in this study. Furthermore, the effects observed depend on the characteristics of the task.
- Key information about eyeglance behavior is not only found in traditionally used metrics like the "Eyes-Off-Road Time Associated With Task Performance" or "Number of Glances to the Task" but also in glances to other locations as well. The analyses demonstrated that very important information is contained in glance durations to the road and mirror locations, for example. This information is

especially relevant for auditory-vocal tasks, which showed different glance patterns than visual-manual tasks, and for event-detection.

- Cognitive distraction effects are very subtle and are not monolithic. Relative to visual distraction, cognitive distraction accounts for much less of the overall variance in driving performance than visual distraction.
- Some effects were observed in the laboratory that were not observed during driving. Until this is better understood, judgments on task effects should reflect a comprehensive evaluation approach that includes more than just laboratory results.

Recommendations for Future Research

Two directions are recommended as next steps for research to be undertaken in the future—naturalistic use studies and skill acquisition, learning, and strategy formulation over time.

Naturalistic Use Studies

The CAMP DWM project was a controlled study in which task performance was requested by the experimenter. It would be of significant value to extend beyond such tightly controlled conditions into an exploration of task performance under naturalistic conditions. This would be undertaken for at least two purposes:

- to determine the extent to which the findings of the DWM project would be replicated under conditions of natural use—particularly with respect to eyeglance patterns and object-and-event detection; and
- to acquire data on the frequency of use for devices, tasks, and activities in vehicles, condition under which specific tasks were initiated, data on strategies of task performance, etc.

Skill Acquisition, Learning, and Strategy Formulation Over Time

Given the conditions of the current study, it was not possible to examine how driver behavior changes as experience with a device increases. This proposed study would examine such questions as:

- whether degradation of driving lessens as skill acquisition improves;
- if learning new strategies for task performance and device usage reduce degradation; and
- how self-paced tests rather than experimenter-paced tests affect driving performance.

Together, these two programs of research would contribute additional knowledge of crash risk with respect to our understanding the complex relationship between the use of in-vehicle devices, workload demands on drivers, degradations in driving performance, and other contributing factors.

Acknowledgements

The Crash Avoidance Metrics Partnership and the Driver Workload Metrics Consortium wish to thank the following individuals for their contributions to the success of this project.

To the core research team, who planned and executed this study, developed the needed instrumentation, analyzed the mountains of data that resulted, dealt with the myriad of details involved with project operations and prepared this report, we are deeply grateful. The long hours and commitment needed were willingly given by the team. Without their contributions, this project would not have been possible. Our heartfelt thanks go to:

Linda S. Angell	James R. Hogsett, Jr.
Jack L. Auflick	Steven M. Kiger
P. Albert Austria	Dev S. Kochhar
Wayne J. Biever	William E. Thomas
Tuhin Diptiman	Louis Tijerina

We also wish to acknowledge the contributions of David L. Smith and Michael Perel, National Highway Traffic Safety Administration, and James P. Foley, James Chang, Daniel Cohen, and Richard Glassco, Mitretek Systems Inc., for their guidance during the planning and execution of this study. Their observations and insights throughout the project were greatly appreciated.

Special thanks go to Dhiraj Tiwari, Mohammed Mustafa, and Sriranga Rao for their efforts in developing the custom software needed to process data from the laboratory and instrumented vehicles, and for their assistance during the analysis phase of the project.

We also greatly appreciate the assistance of Charles Green, Elizabeth Hayes, Adrian Tan, and Marc Winterbottom. These individuals, although only with the project for a brief time, made major contributions nonetheless. The efforts of Adrian Tan and Charles Green in developing the Peripheral Detection Task and in analyzing early pilot testing data must be acknowledged. For their efforts during the initial planning of the study in Task 1, Adrian Tan and Marc Winterbottom should receive recognition, as should Elizabeth Hayes for her similar contributions during Task 2.

Carol Flannagan, Angela Fought, Ian Jolliffe, and Arturo Obscura assisted the project in the area of statistical analysis. We sincerely value their contributions, especially in helping with the interpretation of the analysis results and in providing suggestions on the application of statistical methods.

We are greatly appreciative of the efforts of the drivers, safety observers, experimenters and technicians that participated during the data collection phases of the project. Their dedication through the long hours of training and testing – on weekdays as well as weekends - made it possible to achieve our goals for sample size and completion dates. Our sincere thanks go to the following people:

Stephen Abowd	Heather King
Dorothy Borg	Donna Lew
George Chapp	Brian Mullane
Vanna Cheerala	Ronald Raymer
Paul Ciaverella	Walter Raymer

Robert Edwards	Robert Reeser
Dillon Funkhauser	Michael Robertson
Robert Gadany	Daniel Scott
Rebecca Greedy	Daniel Stuczynski
Roger Hersberger	Linda Wallace
Rhonda Janson	

For their assistance in all of the "back office" activities needed to operate a large-scale study such as this, we are grateful to our administrative support staff. For their efforts in purchasing and procurement support, maintaining project records, entering data into databases, data validation, and the manual verification of events of interest from video recordings, we acknowledge the following for their contributions:

Stephanie Carno	Geneva Kachman
Jennifer Concord	Annette Rashid
Martha Escalante	Roberta Torres
Jill Herbert	Deborah Young

We also wish to thank Dawn Novak, MORPACE International, and her staff for their excellent service in recruiting, screening, scheduling and confirming the test participants for the study. Because of their hard work, the study experienced only a 5 percent "no show" rate across the three data collection phases. This is remarkable considering the two-day commitment required for participation in the study and the remote location of the test track used in Task 2.

Last, but certainly not in the least, we must recognize the efforts by several individuals at The University of Iowa, TNO Human Factors and Noldus Information Technologies USA for their roles in the manual reduction of driver eyeglance from video recordings. Manual reduction of eyeglance data, as those who have conducted this type of work well know, is a long and arduous task. That this work was completed within the time needed and with a high degree of reliability is a testimonial to the expertise and efficiency of all who assisted. The data reduced in this effort represents the sole source of eyeglance data for the project. For his assistance in technical support during the installation of software used to reduce the eye data, we acknowledge Mark Richard, Noldus USA. For their efforts in managing the eyeglance data reduction work and in training the analysts, our thanks go to Daniel McGehee and Mireille Raby of the University of Iowa and to Marieke Martens and Maaike Duistermaat of TNO Human Factors. Finally, the individual analysts, who spent hours in the trenches watching video, must absolutely be commended. Their work was outstanding; their efforts are truly appreciated. These individual included:

<u>The University of Iowa</u>	<u>TNO Human Factors</u>
Timothy Boyle	Maarten Havik
Cheryl Carney	Lien Kalicharan
Carole Simmons	Roland Passau

1 Introduction

1.1 Background

The Crash Avoidance Metrics Partnership (CAMP) was formed between Ford Motor Company and General Motors Corporation in the early 1990s. CAMP's mission is to perform pre-competitive research in areas related to the development of crash avoidance metrics. CAMP provides a mechanism to permit multiple partners in industry and government to cooperate and cost-share research projects that may otherwise be beyond the reach of any single partner to undertake.

The CAMP Driver Workload Metrics (DWM) project was a collaboration between government and industry intended to enhance safety in driving. The DWM project involved the following organizations (presented alphabetically): Ford Motor Company, General Motors Corporation, Nissan Technical Center of North America, Toyota Technical Center-USA, and the United States Department of Transportation (USDOT). The DWM project was funded through the Intelligent Vehicle Initiative (IVI) Light Vehicles Enabling Research Program. It was launched in April 2001 and concluded in March 2005.

1.2 Objectives

The DWM project addressed the measurement of driver distraction related to in-vehicle subsidiary tasks. Generally, in-vehicle device assessments while driving on public roads or test tracks are not feasible early in a product development program. However, early product assessments are necessary because system requirements or designs can be more readily modified and improved early rather than later in the product development process. Therefore, a key goal was to develop or identify repeatable, meaningful, and practical driver distraction metrics.

Some definitions are in order.

- Repeatable means that similar results from one assessment are likely to be found in a comparable assessment using a similar sample of test participants or with different analysts.
- Meaningful means a laboratory metric or analytical model prediction is correlated with at least one aspect of driver performance (lateral control, longitudinal control, and object-and-event detection) or driver eyeglance behavior. Because the laboratory and analytical model metrics are intended to predict aspects of driver performance or eyeglance behavior, they are hereafter referred to as surrogate metrics or measures. (The terms metrics and measures are used interchangeably in this report when referring to laboratory data or analytical model output).
- Practical means that surrogates (either subjective workload assessments, human subjects testing, or analytical modeling) can be implemented within the time and resource constraints typical within the automotive product development environment.

The DWM project was an applied research project rather than a basic or pure research project. The applied objective was to evaluate means with which to readily approximate the distraction potential of in-vehicle tasks in an automotive product development environment. The research procedures largely reflected this applied emphasis. Field testing and mostly realistic tasks were pursued for realism and credibility with original equipment manufacturers (OEMs). A safety emphasis, for relevance, pointed toward measures thought to be related to safety. Usefulness for

OEMs highlighted predictive surrogates that might be feasible to apply in product development. The applied emphasis, rather than basic research, pointed away from refined laboratory testing, abstract tasks of elementary psychological processes, physiological measures that are hard to relate to safety (Chapanis, 1970), etc. A practical emphasis also dictated a focus on first-order task effects. Higher-order effects generally tend to account for much less of the variability observed in performance, even if those higher-order effects are statistically significant (Box, Hunter, and Hunter, 1978). Similarly, individual differences among test participants were a lower-priority topic than in-vehicle task effects on driving. The task was the primary unit of analysis.

Many simplifying assumptions and broad generalizations are presented in this report. Such generalizations, incomplete in their details, may nonetheless be sufficient for applied work (Norman, 1996). Gross generalizations are sometimes presented for brevity and to convey first-order effects that apply across a wide range of conditions (Norman, 1999).

1.3 Driver Workload Defined

The literature indicates that task workload, separate from a participant's abilities, depends on time, task difficulty, and structural interference between concurrent tasks.

Time is fundamental to the concept of workload. A dictionary definition of workload is "the number of hours that a machine, worker, teacher, etc., is required to work in any specific period" (Random House, 1969). Conceptual definitions of "workload as proportional to the ratio of time occupied performing tasks to total time available" also emphasize time as a key component of performance prediction (Wickens and Hollands, 2000).

Task difficulty is often used to describe any task modification that increases the required task time or decreases the accuracy of task completion (Kantowitz, 1987). Sarno and Wickens (1995) point out that more-difficult tasks usually take more time and so will predict greater interference between concurrent tasks. Thus, task difficulty may generally be addressed in reference to time.

"More-difficult tasks take longer" is a useful generalization, but there are exceptions. For instance, difficult tasks might be completed faster than easy tasks if more errors are accepted. This is an example of a speed-accuracy tradeoff (Drury, 1999). Another exception is that long tasks made up of many simple activities may be less demanding than short tasks made up of fewer, more complex activities or processes (cf., Kantowitz, 1985). A third exception is that two tasks of the same duration can have different effects, e.g., just driving for two minutes versus two minutes of destination entry with a complicated route guidance system. Thus, caution should be used in driver workload data interpretation. Several steps may help in that interpretation. Task analysis can provide insights into the nature of the tasks being evaluated. Review of prior research, theory, and modeling also provide guidance. Unsuccessful task performances might be omitted or separated from the analysis of successful task performances. The distraction potential of a task, even if it is a long but monotonous or simple one, might also be assessed in terms of its demands relative to the concurrent demands of driving.

Task duration may be augmented, as needed, by a consideration of structural or resource interference between concurrent tasks (Groeger, 2000; Wickens and Hollands, 2000). The notion of structural or resource interference is based on basic human limitations. Two concurrent visual tasks cannot share foveal vision. Two concurrent auditory-vocal tasks cannot readily share listening and speaking resources. Concurrent tasks that load the same working memory resources can degrade performance on one or both tasks. Resources must be switched from task to task. Less interference is predicted when different input, central processing, or output resources are required by different concurrent tasks. Multiple resource theory is discussed in more detail in the DWM Task 1 report.

Visual-manual tasks and auditory-vocal tasks are fundamentally different in their resource demands from concurrent driving. Driving requires visual inputs to monitor the road scene; spatial working memory to perceive the position, speed, and acceleration of one's own vehicle and others; and manual outputs to adjust steering, accelerator, and brakes. Verbal working memory is also required from time-to-time to read road signs, billboards, bumper stickers, and the like. Visual-manual tasks at a minimum require the same input and output resources as driving. Working memory demands may also overlap. Subsidiary task completion time reflects the duration of resource competition between visual-manual tasks and driving.

Auditory-vocal tasks require auditory inputs, vocal outputs, and (usually) verbal working memory. This implies relatively less structural interference with the driving task. As such, the duration of an auditory-vocal task may have little to do with intrusion on the driving task. A task performed with an auditory-vocal interface may take even longer than the same task with a visual-manual interface and yet it loads the driver less. The lower competition between input-output resources for auditory-vocal tasks and concurrent driving may leave certain aspects of the driving task unperturbed. Working memory demands, on the other hand, may leave at least some aspects of vehicle control unaffected but degrade object and event detection (Brown, 1994). Heightened emotional states can also lead to reduced situational awareness of vehicle control. However, this effect was not addressed in the DWM project.

Concurrent tasks performed while driving may compete with the primary driving task. Tijerina (1996) defined driver workload as the competition between subsidiary tasks and concurrent driving. The driver's primary task is to safely control the vehicle at all times. Safe driving requires the driver to watch the driving scene, steer, manage speed and separation with other vehicles, and detect objects and events in the driving environment in order to respond as appropriate. These aspects of driving define the categories of workload measurement in a driving context.

To summarize the previous points into a definition of workload:

> Workload, in the context of driver distraction, is defined as the competition in driver resources (perceptual, cognitive, physical) between the driving task and a concurrent subsidiary task, occurring over the task's duration, as manifested in degraded lanekeeping, longitudinal control, object-and-event detection, or eye-glance behavior. For the purposes of this research, the workload occurs over the duration of the subsidiary task.

There is no validated transfer function that precisely relates workload measures to crash incidence. Studies that relate selected driver workload measures to crash incidence (e.g., Wierwille and Tijerina, 1998) are best treated as monotonic relations. The basic DWM strategy was to identify and evaluate measures thought to be monotonically related to quality of driving. This means that quality of driving should remain the same or decline as workload increases. The monotone relationship implies that quality of driving should not improve over a practical range as workload increases. This leads to the following relative interpretations when comparing higher-workload to lower-workload in-vehicle tasks:

- More erratic lanekeeping (greater weaving in the lane, more frequent departures out of lane during a task) reflects potentially worse, not better, lateral control while performing a task.
- Greater variation in speed or car following reflects potentially worse, not better, longitudinal control while performing a task.
- More misses reflect potentially worse, not better, object-and-event detection while performing a task.

- More eyes-off-road time reflects potentially worse, not better, driver visual monitoring while performing a task.
- Shorter glimpse times or less frequent glances to the road scene during a visual-manual task reflects potentially worse, not better, visual scanning.
- Fewer mirror checks reflect potentially worse, not better, situational awareness of the surrounding traffic environment.
- Shorter time-to-contact (TTC) reflects potentially worse, not better vehicle separation.

There are other views of the quality-of-driving impacts listed above. These include alternative interpretations such as driver adaptation to in-vehicle task demand, reduction of the driver's quality-of-driving criteria during a task, and others, including the notion that lanekeeping and speed variation may decline due to decreased responsiveness of drivers during short periods of high workload. "Dissociation" from the driving task is also a contrasting point-of-view of workload. This view predicts that high workload might also be manifested in a lack of control inputs or eyeglances. Because of this lack of inputs, dissociation could in fact lead to reduced lane variability, reduced speed variability, increased eyes-on-road time, and so forth. Discussion of these points is deferred until later in this report.

1.4 Project Scope

The DWM project consisted of five tasks:

- **Task 1** set the stage by means of a literature review on measures and methods with which to characterize driver workload; candidate models, simulations, and laboratory metrics and methods that might serve as practical, meaningful, and reliable surrogates for the methods and measures obtained in driving; candidate in-vehicle tasks that span the range of driver demands to which metrics and methods should be responsive; and test scenarios.
- **Task 2** focused on the development of workload metrics and methods through laboratory, on-road, and test track testing. This task also included technical outreach through a series of workshops sponsored by CAMP.
- **Task 3** initially was to validate the practicality, meaningfulness, and reliability of the proposed metrics and methods by use of a new sample of test participants, new tasks, and new evaluators without extensive prior exposure to this project. Task 3 was subsequently withdrawn by mutual agreement between USDOT and CAMP.
- **Task 4** focused on project documentation.
- **Task 5** encompassed project management tasks.

1.5 Scope Exclusions

The DWM project addressed driver workload issues. It aimed to develop measures of in-vehicle task demands while driving. This placed an emphasis on the negative impacts of in-vehicle tasks on driving rather than the negative effects of driving on in-vehicle tasks. There were other aspects of driver workload considered outside the scope of activities in this project. These included:

- **Design guidelines:** Task workload can be addressed through task and equipment design, but design guidelines were outside the scope of the project.
- **Human abilities:** Task workload differs for people with a range of human abilities, but this variation is taken as a given among the automotive customer base. Human abilities may be addressed by training or selection, but these are incidental to the DWM project.
- **Emotionally-laden tasks:** Also outside of the scope of this project were potentially emotion-laden tasks such as cell phone conversations on politics, religion, current affairs, or significant life events (e.g., Strayer, Drews, Albert, and Johnston, 2002; Drews, Pasupathi, and Strayer, 2004). The tasks selected for study in this project did not use emotionally-laden content.
- **Driver underload:** Workload also has a facet called underload, i.e., insufficient activity to maintain normal performance. Underload was also considered outside the scope of this project.
- **Exposure factors:** The DWM project also did not address exposure factors. There was no attempt to catalogue exposure factors such as frequency of use, road location of use, time-of-day of use, season of use, driver traits, driver states, driver support systems (e.g., collision warning systems) during use, task content, device implementation, and so forth.

1.6 Report Organization

This report presents the empirical results of Task 2 of the DWM project. It includes an overview of the data collection strategy used in this research. Results from public road and test track trials are presented with an emphasis on the task effects on driving performance and driver eyeglance behavior. Findings for a variety of laboratory surrogate methods and metrics are presented. These include laboratory methods that involve human subjects testing as well as subjective workload assessments. In addition, an analytical model that produced a variety of workload estimates was developed and evaluated. The project also examined the impact of driver individual differences on workload measures. The report concludes with a discussion of the overall pattern of results, recommendations of the methods of driver workload assessment, and recommendations for future research. Appendices provide greater detail on the procedures, materials, and data presented.

1.7 Chapter References

Brown, I. D. (1994). Driver fatigue. *Human Factors*, 36(2), 298-314.

Box, G. E. P., Hunter, W. S, and Hunter, J. S. (1978). *Statistics for experimenters*. New York: John Wiley and Sons.

Chapanis, A. (1970). Plenary discussion: Relevance of physiological and psychological criteria to man-machine systems: The present state of the art. *Ergonomics, 13(2), 337-346.*

Drews, F., Pasupathi, M., and Strayer, D. L. (2004). Passenger and cell phone conversations in simulated driving. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting.*

Drury, C. G. (1999). Managing the speed-accuracy trade-off. In. W. Karwowski and W. S. Marras (Eds.), *Occupational ergonomics handbook* (pp. 677-691). Boca Raton, FL: CRC Press.

Groeger, J. (2000). *Understanding driving: Applying cognitive psychology to a complex everyday task*. London: Taylor and Francis.

Kantowitz, B. H. (1985). Stages and channels in human information processing: A limited analysis of theory and methodology. *Journal of Mathematical Psychology, 29*, 135-174.

Kantowitz, B. H. (1987). Mental workload. In P. A. Hancock (Ed.), *Human factors psychology* (pp. 81-121). Amsterdam: Elsevier Science Publishers B. V. (North-Holland).

Norman, D. A. (1996). The post disciplinary revolution: Industrial design and human factors – heal yourselves. Combined keynote address to Industrial Design Society of America and 1996 Annual Meeting of the Human Factors and Ergonomics Society.

Norman, D. A. (1999). *The invisible computer: Why good products can fail, the personal computer is so complex, and information appliances are the solution* (pp. 193-194). Cambridge: MA: MIT Press

Random House. (1969). *Random House Dictionary of the English language: College edition* (L. Urdang, Ed.). New York: Random House.

Sarno, K. and Wickens, C. D. (1995). Role of multiple resources in predicting time-sharing efficiency: Evaluation of three workload models in a multiple-task setting. *International Journal of Aviation Psychology, 5*(1), 107-130.

Strayer, D., Drews, F., Albert, R. W., and Johnston, W. A. (2002). Why do cell phone conversations interfere with driving? *Proceedings of the 81st Annual Meeting of the Transportation Research Board*, Washington, DC.

Tijerina, L. (October, 1996). *Final Report – Program executive summary: Heavy vehicle driver workload assessment* (Report No. DOT HS 808 467). Washington, DC: National Highway Traffic Safety Administration.

Wickens, C. D., and Hollands, J. G. (2000). *Engineering psychology and human performance* (Third edition). Upper Saddle River, NJ: Prentice Hall.

Wierwille, W. W., and Tijerina, L. (1998). Modeling the relationship between driver in-vehicle visual demands and accident occurrence. In A. G. Gale, I. D., Brown, C. M. Haslegrave, and S. P. Taylor (Eds.), *Vision in Vehicles-VI* (pp. 232-244). Oxford: North-Holland.

2 Study Design Overview

2.1 Introduction to Venues

This chapter provides a high-level summary of the data collection strategy used in this research. Details are provided in the appendices.

A different sample of participants was recruited for each of three test venues: laboratory testing, on-road testing, and test track testing. The participants in the on-road testing were different than those who participated in the test track testing. The participants who volunteered for the laboratory testing were different from those who participated in either the on-road testing or test track testing. However, each participant sample was obtained based on the same screening criteria. This plan provided for between-participants comparisons across venues to assess the predictive validity (i.e., meaningfulness) of laboratory surrogates and the comparability between road and track results.

Test participants were asked to perform a set of requested tasks while their performance was recorded (see Table 2-1). The test plan scheduled each participant to perform each of the requested tasks assigned for a particular test method. This plan provided repeated measures across tasks for more sensitive testing. Thus, the basic data collection plan involved a two-way, participants-by-tasks layout.

Each participant was scheduled to perform each task twice under nominally the same conditions. This plan provided replications (reps) to assess the repeatability of the various measures collected in the study. The reps also could be combined to create a single measure per participant. Combining data in this manner can make more data available on a per subject basis when one of the subject's replications is missing.

Three separate studies were conducted with data collected in each of three test venues:

- Data was collected for a subset of requested tasks while participants drove on public roads outside of Detroit.
- Data was collected for the on-road tasks plus additional, more demanding tasks while participants drove on the high-speed test track of the Ford Motor Company Michigan Proving Ground.
- Data was collected on all tasks using different laboratory (surrogate) methods in the CAMP Driver Workload Metrics laboratory located in Farmington Hills, Michigan.

Tasks were generally ordered (1^{st}, 2^{nd}, 3^{rd}, 4^{th}, etc.) for presentation according to a diagram-balanced, Latin square (Lewis, 1988). This counterbalancing approach ensured that tasks appeared an approximately equal number of times in each serial position across the sample of participants for a particular test. Such a procedure is used to control for nuisance variables like fatigue, learning effects, and driving conditions. One exception was the CD/Track 7 task. This task was added late in the on-road test execution and was assigned to the last blocks of the day.

Tasks were scheduled in blocks comprised of sets of tasks. Within a block of tasks, each task was performed once. In a subsequent replication block, each task was again performed once. A given task sequence from the Latin square was assigned to a test participant and this task sequence was used for all blocks.

Each test participant volunteered for two sequential days of testing, regardless of venue. Details on individual venues are provided below.

Table 2-1. Resource Requirements for 23 Requested Tasks

		INPUT MODE		WORKING MEMORY		OUTPUT MODALITIES	
TASK	**DESCRIPTION**	**Visual**	**Auditory**	**Spatial**	**Verbal**	**Manual**	**Vocal**
Just Drive	Baseline task of driving alone	✓		✓		✓	
HVAC	Adjust fan speed, temperature, and vents	✓		✓		✓	
Insert Cassette	Take a cassette from its case and insert the specified side (Side A or B) into player	✓		✓		✓	
Coins	Select specific amount (e.g., 65¢) from coins in a center-console cup holder	✓		✓		✓	
Radio (Easy)	Tune to a specified frequency given via MP3 recording; initially, radio is ON, on the appropriate frequency band, at a preset frequency	✓			✓	✓	
Radio (Hard)	Tune to a specified frequency given via MP3 recording; initially, radio is OFF, on the wrong frequency band, at a preset frequency	✓			✓	✓	
CD / Track 7	Take a specified color-coded CD (e.g., the red CD) from visor-wallet, insert into car radio-unit, select Track 7	✓		✓ (for color)	✓ (to read track #)	✓	
Manual Dial - Home	Enter own area code and home phone number into a flip phone	✓			✓	✓	
Voice Dial	After pressing <*> <TALK> keys, voice-dial home phone number using 10 digits	✓ (to start)	✓		✓	✓ (Initial keying then hold)	✓
Destination Entry	Enter street address (city, state, street name, and number) into Magellan navigation system	✓			✓	✓	
Route Tracing	Trace a route through a paper maze from a point of origin to a point of destination	✓		✓		✓	
Read (Easy)	Silently read ~30 word narrative printed on paper at 4th to 5th grade Flesch-Kincaid Reading Score Level, say aloud missing word at end	✓			✓		✓

		INPUT MODE		WORKING MEMORY		OUTPUT MODALITIES	
TASK	**DESCRIPTION**	**Visual**	**Auditory**	**Spatial**	**Verbal**	**Manual**	**Vocal**
Read (Hard)	Silently read ~60 word narrative printed on paper at 7th to 8th grade Flesch-Kincaid Reading Score Level, say aloud missing word at end	✓			✓		✓
Map (Easy)	Say aloud relative orientation of two destinations on a paper map with 12 such destinations in call-out boxes	✓		✓	✓ (minor)		✓
Map (Hard)	Say aloud relative orientation of two destinations on a paper map with 22 such destinations in call-out boxes	✓		✓	✓ (minor)		✓
Travel Computations	Mentally compute and say aloud distance traveled, toll sums, time to arrival, fuel needed from information presented via audio messages		✓		✓		✓
Sports Broadcast	Listen for who a requested team (e.g., the Phillies) played and who won from a broadcast covering many teams, then reply		✓		✓		✓
Book-on-Tape Listen	Listen to a recording of a short (2-minute) mystery story to subsequently summarize the story line		✓		✓		✓
Book-on-Tape Summarize	Orally summarize the Book-on-Tape Selection				✓		✓
Biographical Q&A	Listen to and orally reply to simple biographical questions (e.g., name, address, type of car driven, etc.)		✓		✓		✓
Route Instructions	Listen to and repeat back recorded routing instructions		✓	✓	✓		✓
Route Orientation	Listen to recordings of direction-of-travel and subsequent turn, then answer "what direction are you traveling now?"		✓	✓			✓
Delta Flight Information (Delta Flightline)	After voice-dial, seek out arrival time for an origin-destination city pair with a given departure time	✓ (to start)	✓		✓	✓ (Initial keying then hold)	✓

Note: Voice Dial and Delta Flight Information tasks were mixed-mode tasks. Each task began by picking up a cell phone and pressing a short key sequence (e.g., <*><TALK> for Voice Dial). Thereafter, the task was primarily auditory-vocal in nature, though the cell phone was held until the task was completed.

2.1.1 On-Road Venue

A subset of the DWM tasks was evaluated on public roads. These tasks were chosen for safety considerations. They were predicted to be less demanding than additional tasks reserved for track testing alone. On-road data collection used a three-car platoon. Testing was planned for daylight driving on dry pavement. The test participant drove a subject vehicle equipped for data collection. A lead vehicle ahead of the subject vehicle presented a consistent car-following stimulus and a lookout for traffic situations ahead of the platoon. A follow-vehicle, about 2 to 3 car lengths behind the subject vehicle, provided a consistent following-vehicle stimulus and a lookout for surrounding traffic coming up from the rear.

Four CAMP project staffers accompanied each test participant for the on-road testing. A safety observer sat in the front passenger seat of the subject vehicle. The safety observer had responsibility to ensure the participant did not lose control of the vehicle. An experimenter sat in the rear passenger seat of the subject vehicle. The experimenter orchestrated the schedule and presentation of task requests, the launch of object and event detection (OED) stimuli (described below), and the monitoring of the data acquisition system. The follow vehicle driver had call-off authority for object-and-event detection stimulus events if surrounding traffic precluded it. The lead-vehicle driver was responsible for assessing the traffic ahead, and had call-off authority for lead-vehicle deceleration events (described below) if driving conditions precluded it. The lead-car driver was responsible for maintaining a constant speed (through cruise control) for car following and staying above a minimum speed and separation distance for deceleration events.

Each test participant was required to complete the on-road testing over a period of two days. Participant intake (informed consent, familiarization with the testing protocol), and task training occurred the morning of Day 1. Individual differences testing was conducted either in the morning of Day 1 or the afternoon of Day 2, depending on the time available. Testing on public roads began after lunch on Day 1 and continued through Day 2. Participant debriefing during the afternoon of Day 2 concluded the two-day session.

After lunch on Day 1, the participant was familiarized with the subject vehicle and its controls. A five-minute vehicle familiarization video was shown to each participant to better acquaint the participant with the controls and operation of the vehicle. The participant also was shown the lead-vehicle to be followed at all times and the follow-vehicle that stayed behind the subject vehicle during on-road testing. Safe driving was emphasized and the participant was made aware that there was no obligation to perform a task if the participant felt uncomfortable with the driving conditions. Video calibration was performed every time the participant re-entered the vehicle.

The east- and westbound lanes of Interstate I-96 between Brighton (Exit 145) and Williamston (Exit 117), Michigan, were chosen as the test route due to relatively low traffic volume and the evenness of the road geometry and posted speeds. The participant was asked to drive in the right lane at all times and at about 55 mph during the testing.

The participant drove the subject vehicle for approximately 30 minutes to commute to the test route. This provided the participant with an opportunity to become familiar with the vehicle's feel and controls. The experimenter explained what the participant was required to do and the objects and events they were to monitor during the tests trials. Three OED stimuli were chosen for this study. The lead vehicle center high-mounted stoplight would sometimes turn on for a duration equal to the instantaneous time headway at stimulus onset (i.e., inter-vehicle distance divided by subject vehicle speed at onset). The driver-side follow-vehicle turn signal (FVTS) would sometimes illuminate for a fixed 2.5 seconds. Sometimes, lead-vehicle deceleration (LVD) would occur when the lead-vehicle driver received an instruction to disengage cruise control and coast down (no braking or brake lights) from a nominal 55 mph to no less than 45 mph or until notified

over two-way radios to surge ahead. The test participant was trained to resume normal driving after detecting the LVD event and close the gap between the subject and lead vehicle to 3 to 5 car lengths. The test participant responded to the CHMSL and the FVTS stimuli, if detected during task performance, by pressing a button attached by a hook and loop closure to the left index finger. To respond to the LVD event, the participant was required to gently tap the brakes to indicate event detection. OED stimuli were randomly assigned within a sequence of tasks in a block. In the replication for that block, the same OED assignments applied. Long-duration (e.g., longer than 60 seconds) tasks were assigned three OED stimuli per trial. These stimuli were scheduled to be presented during the long task's duration. The order of OED stimuli was permuted across blocks of trials. The experimenter explained each of the three OEDs and provided examples for the benefit of the participant. Participants also were informed that they were to attempt to follow about three to five car lengths (approximately 120 feet), behind the lead-vehicle during the task. The experimenter, with the assistance of the radar installed in the car, coached the participant into the appropriate following distance. Between trials, the experimenter directed the test participant to speed up or fall back as needed to prepare for the next task.

Eight blocks of data collection were scheduled through the afternoon of Day 1 and the morning of Day 2. The first six blocks of data per participant had to be completed to meet the counterbalanced task layout with respect to the object-and-event detection stimuli. Blocks 7 and 8 provided trials for additional tasks identified immediately prior to the start of testing (e.g., CD/Track (7). These two blocks also provided single OED stimulus presentations for selected long duration tasks. Replicates (same driver, vehicle, road, task, and OED events) were in adjacent blocks (1 and 2 were replicates, 3 and 4 were replicates, etc.).

Every effort was made by each experimenter to complete the testing as designed. However, many factors contributed to missed trials. These factors included traffic conditions, weather, hardware or software problems, procedural errors, time constraints, etc.

After lunch on Day 2, the participant returned with the experimenter and staff to CAMP where additional Day 2 formalities were completed as time permitted. Participants were asked to complete questionnaires of multitasking difficulty, overall workload, and situational awareness. In addition, each participant was required to complete two blocks of the static task time and occlusion surrogates if time permitted. If the participant was unable to complete any of the individual differences testing activities on Day 1, they were completed in the afternoon of Day 2. The participant was required to sign an end-of-Day-2 data collection form prior to being released from the study.

2.1.2 Test Track Venue

Track testing was conducted on the high-speed test track of Ford Motor Company’s Michigan Proving Ground in Romeo, Michigan. The test track was a 5-mile oval with 1-mile straightaways and 2,500-foot radius curves. The track has 5 lanes, each 12 feet wide on the straightaways that transition to 13 feet wide on the curve sections. The three-car platoon was given access to lane 2, which was unbanked on the curve sections.

Procedures used for on-road testing were applied to track testing as well. Two laboratory stations were set up at the proving grounds for task training and individual differences testing purposes. No surrogate data was collected at the test track. The participants were greeted at the lobby at 7 a.m. and were escorted to the testing site by the experimenters. The order of protocols was similar to that of the on-road segment; however, in this segment all 23 DWM tasks were used in the data collection process, compared to the 16 that were used in the on-road testing. Sufficient time was provided for the participant to drive around the track and get a feel for the vehicle dynamics and the controls.

Track testing began after lunch and continued through the morning of Day 2. There were six blocks of tasks for track testing. Each block contained all 23 tasks. Each task was assigned one and only one OED stimulus per trial. The OEDs and the testing protocols used in the test track were identical to the ones used on the on-road segment. The OEDs were explained to the participant and examples were provided during the practice drive. Track trials were conducted as the final venue for data collection. Experience during the on-road trials suggested certain procedural efficiencies that could be applied to the track venue. Procedural simplification was in three areas. Track testing would use only six blocks, three to cover all tasks and three blocks for replication. Each trial would have one and only one OED event scheduled for it. The track testing would schedule all three replication blocks only after all three blocks of original blocks were completed. The reduced number of blocks and revised OED schedule were more readily executed on the track than the eight blocks used for the on-road trials. The notion of scheduling replication blocks only after all tasks had been tested once was due to experiences with weather. The goal was to have at least one trial per task per OED scheduled in the event that rain cut short the testing session. Road testing had replications in blocks back-to-back. A 10-minute break was provided after each block in addition to a one-hour lunch break.

Day 2 testing for the participant was the same as the on-road venue.

2.1.3 Laboratory Venue

Laboratory testing examined both performance testing and subjective workload assessments. Six different surrogate methods for performance testing of tasks were investigated in the laboratory. Tasks were assessed twice with each method. Brief descriptions of these methods are provided below and details of the methods are provided in Appendices D and R.

Static Time This method provided a static time metric. Measurements were taken of the total time needed to complete each variable-duration visual-manual DWM task when performed alone, without any concurrent task or any interruptions

Occlusion This method provided a total shutter open time (TSOT) metric. Measurements were taken during task performance of the TSOT needed to complete a visual-manual DWM task when performed wearing occlusion goggles. These goggles were computer-controlled to open for 1.5 seconds and close (go opaque) for 2.0 seconds cyclically until the task was done.

PDT-Alone This method provided percent missed detections and reaction time metrics. Measurements were taken during DWM task performance of the number of missed detections and the reaction time of detections to a peripheral detection task (PDT) light. The PDT light was a high-intensity spot of laser light. It was briefly and periodically presented on a wall-mounted projection screen in front of the test participant during task performance. The participant pressed a button with the left hand if the PDT light was noticed. Multiple PDT stimuli were presented during task performance, more for longer tasks. All DWM tasks were assessed.

PDT with STISIM

This test provided task duration, lanekeeping, and speed maintenance measures while driving and performing DWM tasks. Measurements of driving performance in the Systems Technology Inc. (STI) fixed-base, part-task simulator were taken as subjects concurrently drove the simulator and performed requested tasks. The simulation involved following a lead vehicle that traveled at constant velocity at a consistent self-selected following distance. The PDT stimuli described above were also concurrently presented during the drive. All DWM tasks were assessed.

Sternberg-Visual

This method provided percent error and reaction time metrics, as well as percent missed detections. A participant memorized three symbolic road signs (e.g., T-intersection, traffic merging from right, road enters from left, etc.) prior to the start of a trial. During CAMP task performance, individual symbolic road signs were periodically and briefly presented on an LCD display mounted ahead of the participant. When noticed, the participant was instructed to press one button with the left hand if the presented road sign was from the memorized set (a positive-match probe) or different button with the left hand if the presented road sign was not from the memorized set (a negative-match probe). Multiple probe signs were presented sequentially on the LCD display, more for longer tasks. Approximately equal numbers of positive and negative probes were scheduled for presentation in a random order. The road signs for the Sternberg-Spatial test were of road geometry and did not involve alphanumerics. All DWM tasks were assessed.

Sternberg-Verbal

This test was identical to the Sternberg-Spatial test except the memorized signs and subsequent probes were of alphanumeric signs like state route numbers. All DWM tasks were assessed.

Participants also completed a variety of subjective assessments of tasks:

Operator Workload (OWL)

A univariate scale of overall task workload from 1 (low) to 100 (high). The participant rated a given task against this fixed scale. All tasks were assessed except the Just Drive task.

Magnitude Estimation of Multitasking Difficulty

A rating scale of how hard it is to do a given task while driving and maintaining lane position, speed, headway, and detecting objects and events on or near the roadway. Ratings were elicited with a standard task of turning on the radio, switching to the FM band, and tuning to a specific frequency. The test participant was presented a standard for comparison (radio tuning task) and this standard was arbitrarily assigned a scale value of 100 by the experimenter. The test participant was asked to make judgments about other stimuli (e.g., each other task), and reflect how many times greater or lesser one stimulus might be as compared to the standard. That is, the participant was asked to estimate the ratio between the two stimuli, sensations, perceptions, or judgments. If a stimulus seemed twice as great as the standard, the test participant should say "200." If a stimulus appeared only half as great as the standard, the test participant should say "50" and so on. All tasks were assessed except the Radio (Hard) tuning task (used as the standard for comparison) and the Just Drive task.

Magnitude Estimation of Situational Awareness

This is a rating scale of how aware test participants felt they would be to the roadway traffic and events while performing each task as compared to the standard task of turning on the radio, switching to the FM band, and tuning to a specific frequency. All other details are similar to the magnitude estimation for multitasking difficulty.

Each test participant completed performance testing in each surrogate twice to provide two replications for analysis. The subjective assessments were also completed twice except for the Situation Awareness ratings, which were collected only once. In addition to the counterbalancing of tasks evaluated within a given surrogate method, the order of surrogates was counterbalanced.

The focus of this study was on the effects of tasks on performance. Differences among individuals can contribute to differences in performance while performing tasks. This was a secondary research interest that was supported by collection of demographic information about the participants (age, gender, high-technology device familiarity, self ratings of multitasking ability, etc.). Selected tests of human abilities were also administered. Computerized tests included the Useful Field of View test, and the PATSYS versions of Manikin, Dynamic Visual Acuity, and Baddeley Grammatical Reasoning tests (Tijerina, Parmer, and Goodman, 1999). In addition the Baddeley Dual-Task paper-and-pencil test was administered (Della Sala, Baddeley, Papagno, and Spinnler, 1995). These tests are described in Chapter 7, *Individual Differences and Driver Workload Metrics*.

The laboratory testing was conducted over two days in a modified office space with four testing stations, each fully equipped for testing. The morning of Day 1 focused on participant intake, task training, and surrogate test familiarization. A 10-minute break was provided for every hour of testing for both days of testing. After lunch, participants resumed task training and surrogate training activities as needed prior to beginning actual testing sessions. Individual differences tests were conducted during the afternoon session. Participants were asked to sign a Day 1 completion form and arrive at 8 a.m. the next day for the remainder of the study.

Upon the arrival of the participant on Day 2, the experimenters proceeded to complete the remainder of the surrogate testing. After completing surrogate testing, individual differences tests were conducted if they were not completed on Day 1 of testing. The participants were then asked to complete the overall workload and multitasking difficulty ratings prior to signing a Day 2

completion form and being released from the study. In addition to the multitasking difficulty rating, a situational awareness rating scale was introduced for Day 2.

2.2 Tasks Evaluated in Each Venue

Table 2-2 indicates what tasks were evaluated in each venue. Fewer tasks were evaluated in the on-road venue because higher-demand tasks were omitted. The reduced traffic of the test track venue allowed all tasks to be evaluated. Different laboratory tests and ratings were applicable to different sets of tasks. For example, TSOT and the R-Metric (the ratio of TSOT to Static Time) only applied to visual-manual tasks. The static time method could not be applied to study tasks of fixed duration. The subjective ratings of workload, situation awareness, and multitasking difficulty omitted a Just Drive condition because it was not dual-task. Additionally, the Radio (Hard) task was arbitrarily assigned a value of 100 by the experimenter. It was not rated by test participants because it was the standard for comparison needed for magnitude estimation. Finally, some laboratory methods like PDT Alone and Sternberg allowed for data collection of performance on just that method alone. For convenience, only the phrase “Just Drive” is used though its meaning should be clear from context.

Table 2-2. Tasks Evaluated in Each Venue

TASK	TRACK TEST	ROAD TEST	LAB: SUBJECTIVE ASSESSMENTS			LAB: PERFORMANCE MEASURES						
			OWL	Sit. Aware.	Multi-Task	Static Time	TSOT	R-Metric	PDT Alone	PDT & STISIM	STISIM	Sternberg
Just Drive	✓	✓							Just PDT	✓	✓	Just Sternberg
HVAC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Insert Cassette	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Coins	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Radio (Easy)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Radio (Hard)	✓	✓	✓	Modulus, or standard task to which others were compared	Modulus, or standard task to which others were compared	✓	✓	✓	✓	✓	✓	✓
CD / Track 7	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Manual Dial - Home	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Voice Dial	✓	✓	✓	✓	✓				✓	✓	✓	✓
Destination Entry	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Route Tracing	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Read (Easy)	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Read (Hard)	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Map (Easy)	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Map (Hard)	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

TASK	TRACK TEST	ROAD TEST	LAB: SUBJECTIVE ASSESSMENTS			LAB: PERFORMANCE MEASURES						
			OWL	Sit. Aware.	Multi-Task	Static Time	TSOT	R-Metric	PDT Alone	PDT & STISIM	STISIM	Sternberg
Travel Computations	✓	✓	✓	✓	✓				✓	✓	✓	✓
Sports Broadcast	✓	✓	✓	✓	✓				✓	✓	✓	✓
Book-on-Tape Listen	✓	✓	✓	✓	✓				✓	✓	✓	✓
Book-on-Tape Summarize	✓	✓	✓	✓	✓				✓	✓	✓	✓
Biographical Q&A	✓	✓	✓	✓	✓				✓	✓	✓	✓
Route Instructions	✓	✓	✓	✓	✓				✓	✓	✓	✓
Route Orientation	✓	✓	✓	✓	✓				✓	✓	✓	✓
Delta Flight Information	✓		✓	✓	✓				✓	✓	✓	✓

2.3 Prior Predictions of Relative Task Workload

Known demand effects of the DWM tasks on real-world driving impacts would help interpret the research results. Such data are generally not available. The most direct data would presumably be contained in Police Accident Reports (PARs). But PARs do not identify crash causal factors definitively except in those rare cases where physical evidence exists (a hand-held phone open and on the floor board, a dash-mounted DVD player on and playing, etc.) and the investigating officer notes it. Crash databases (Wang, Knipling, and Goodman, 1996; Stutts, Reinfurt, Staplin, and Rodgman, 2001) also do not provide such data, in part because task demand and exposure are intertwined. Furthermore, it is not known to what extent drivers report that they were distracted when other factors were actually involved. Epidemiological studies (e.g., Violanti and Marshall, 1996; Redelmeier and Tibshirani, 1997; Laberge-Nadeau et al., 2003) support plausible inferences but cannot yet provide definitive evidence of key variables. For example, epidemiological studies generally cannot determine the exact moment of crash impact but must estimate it instead (Redelmeier and Tibshirani, 2003).

Task workload predictions, independent of the project data, are needed to aid interpretation of the results and avoid circular reasoning. To solve this problem, human factors literature, human performance theory, and models of dual-task workload were used to provide these prior predictions. These sources were used to provide workload demand predictions that identify relative task workload into lower-workload and higher-workload task categories.

A coarse categorization of relative task workload prediction into only two levels of demand is considered appropriate for several reasons:

- The current state of the art of driver workload prediction does not allow for fine gradations in task-related workload.
- Original Equipment Manufacturer (OEM) decision making is often of an acceptance testing variety (acceptable versus not-acceptable).
- This approach helps manage the post-hoc paired comparisons problem. Without any limitations, 23 tasks would require (23*(23-1))/2 or 253 paired comparisons per measure. The number of paired comparisons is substantially reduced if only tasks across categories (rather than within a category) are compared.
- Further improvement in statistical power is gained because directional or one-tailed tests can be used. For example, if measures are oriented toward a "more implies more workload" order, then directional tests can be made of the general null and alternative hypotheses (one could reverse the sign of the alternative hypothesis for a less is more workload-oriented measure):
  - Ho: higher-workload task result = lower workload task result
  - Ha: higher-workload task result > lower workload task result.

2.3.1 Basis of Higher-Workload and Lower-Workload Categorization

Prior prediction depended on sorting DWM tasks into higher-workload and lower-workload categories. Three prediction sources were used to do this. Research and theory in human factors and cognitive psychology was one source of guidance. CAMP DWM analytical modeling of task workload was a second source of guidance. The engineering judgment of the DWM investigator who developed the approach was a third source of guidance. The majority of predictions across the three sources resulted in the final categorization.

Prior predictions of relative workload were based on several sources of literature, theory, and modeling. These included application of Multiple Resource Theory (MRT) (Wickens and Hollands, 2000; Groeger, 2000), modified MRT modeling, task time models (Card, Moran, and Newell, 1983; Harris, Iavecchia, and Bittner, 1988; Nowakowski and Green, 2001), and content analysis of specific tasks.

Table 2-1 provides a depiction of the 23 DWM tasks in terms of task dimensions from MRT (Wickens and Hollands, 2000; Groeger, 2000). This theory predicts that tasks that require the same resources will interfere more than tasks that require different resources. The driving task used in the DWM project required primarily visual input to monitor the driving scene; spatial working memory to process lane position, speed, and vehicle separation; and manual output to use the steering wheel and pedals. MRT predicts that visual or visual-manual tasks with spatial working memory requirements should interfere more with driving than auditory or auditory-vocal tasks that use primarily verbal working memory. A modified MRT model was used to make theory-based predictions of DWM tasks in Table 2-1. At a most basic level, visual or visual-manual tasks would be separated from primarily auditory or auditory-vocal tasks.

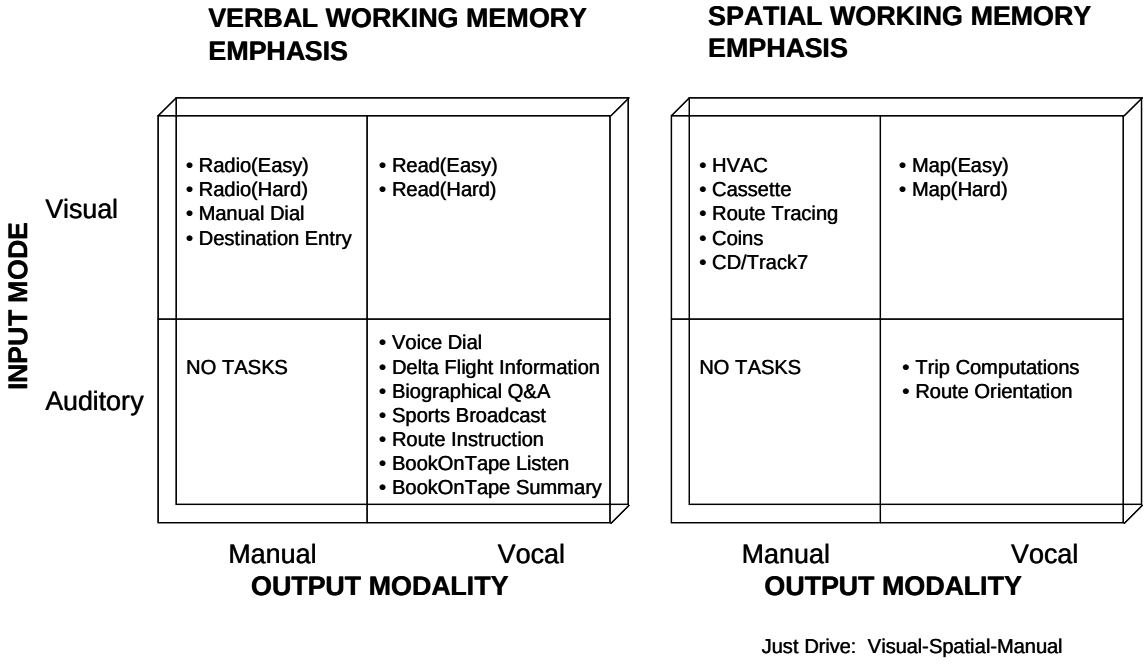


Figure 2-1. Multiple Resources Dimensions of DWM Tasks

Time is a critical component of workload that can be used to predict relative workload within the set of visual-manual or auditory-vocal tasks. An Activity Time Model was developed in this project to estimate task durations. This model provided predicted task durations independent of the CAMP DWM data. The Activity Time Model used activity time prediction equations derived from the work-sampling literature (Niebel, 1976; Smith, 1978) and the cognitive science literature (Card, Moran, and Newell, 1983; Harris, Iavecchia, and Bittner, 1988; John and Gray 2000; Laughery Archer, and Corker, 2001). These equations and values were applied to task steps identified by detailed task analysis of the study tasks.

Figure 2-2 presents the categorization of CAMP DWM tasks into relative lower-workload and higher-workload groups. This is done separately for visual-manual tasks and auditory-vocal tasks. The Just Drive task was included with the auditory-vocal tasks for reasons explained later in this chapter. Mixed-mode tasks like Voice Dial and Delta Flight Information (Delta Flightline) were not categorized because they are not as well understood, specifically with regard to voice recognition system performance impacts, and they have characteristics of both visual-manual and auditory-vocal task types. The basis of this figure is described below.

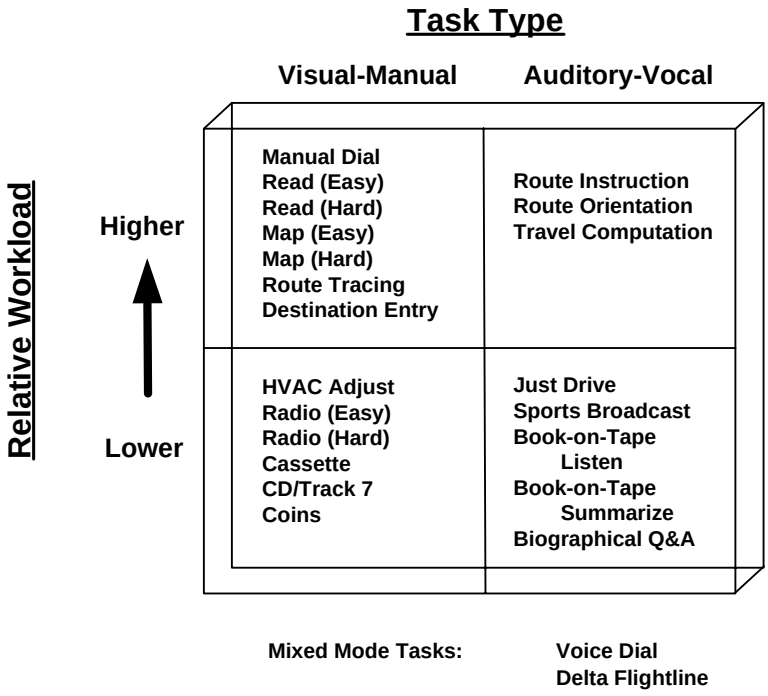


Figure 2-2. Prior Predictions for Relative Task Workload

2.3.2 Visual-Manual Tasks

A rule was needed to categorize the DWM visual-manual tasks into higher-workload versus lower-workload categories. Manual Dial was chosen as the defining task. Per prediction source, a task was chosen for inclusion in the higher-workload category if its estimated demand was as high as or higher than the Manual Dial task. If a task's estimated demand was less than Manual Dial, it was categorized as a lower-workload task.

The Manual Dial categorization rule was based on a limited review of the literature. The DWM task of Manual Dial was chosen, in part, because general trends in existing research indicate this task generally (though not always) imposes greater workload than conventional tasks such as HVAC adjustment or radio tuning (Goodman, Tijerina, Bents, and Wierwille, 1999; references in Goodman, Barker, and Monk, 2005).

A second prediction source was based on CAMP analytical modeling—specifically Activity Time estimates. Activity Times were generated from the CAMP analytical model for workload prediction. The DWM model used micro-models of activity times based on data that was outside of the measures taken in the CAMP DWM project. These micro-models were applied to each task step identified through task analysis. This model is discussed in greater detail in Chapter 6, *Analytical Results*. CAMP DWM tasks with estimated median (i.e., across task analysts) activity times less than that of Manual Dial were categorized as lower-workload tasks.

DWM task analysts also characterized each task step of each task along the dimensions of MRT. The CAMP analytic model generated Dual Task Conflict Potential (DTCP) values for each task. The DTCP results showed anomalies of differences with respect to existing research and task content. For example, the median-of-analysts DTCP values showed higher DTCP values for Manual Dial-Home and HVAC than those for Destination Entry. It should be noted that the Modified MRT predictions derived from the initial development of a model based on MRT were not anomalous with respect to HVAC, Manual Dial and Destination Entry. However, when this model was taken further, in an attempt to provide finer discrimination between tasks, these anomalous results began to emerge. It is the refined model that generates DTCP. It is for this reason that the model output was not included. The more detailed predictions were required, and did not match prior expectations.

The third prediction source was the engineering judgment of the investigator who developed the prior prediction approach for this project. These judgments reflect familiarity with the area of driver distraction and research experience in the field. It also reflects the expert's approach to research in the area of driver distraction.

Table 2-3 presents the basis for prior prediction into lower-workload and higher-workload categories for the visual-manual tasks.

Table 2-3. Basis of Relative Higher Versus Lower Workload Prior Prediction for Visual-Manual Tasks

Task	Prior Research	Activity Time Modeling-Median of Raters	Engineering Judgment	Final Prediction
Coins	L	H	L	Lower
Insert Cassette	L	L	L	Lower
HVAC	L	L	L	Lower
Radio (Easy)	L	L	L	Lower
Radio (Hard)	L	L	L	Lower
CD/Track 7	L	H	L	Lower
Manual Dial - Home	H	H	H	Higher
Route Tracing	H	H	H	Higher
Destination Entry	H	H	H	Higher
Read (Easy)	H	H	H	Higher
Read (Hard)	H	H	H	Higher
Map (Easy)	H	H	H	Higher
Map (Hard)	H	H	H	Higher

2.3.3 Auditory-Vocal Tasks

Auditory-vocal tasks were harder than the visual-manual tasks to categorize into higher-workload and lower-workload sets. An attempt was made to use the same three prediction sources, i.e., prior research trends, CAMP analytical modeling, and engineering judgment. In the end, these three sources were not independent and engineering judgment played a larger role in the task categorization than it did for the visual-manual tasks.

The categorization rule to sort tasks was based on prior research and theory, as well as task content. Rather than a single task, the investigator categorized Route Orientation, Route Instructions, and Travel Computations as higher-workload tasks. The remaining auditory-vocal tasks were assigned to the lower-workload category, as was the Just Drive Task. Just Drive was included in the auditory-vocal task set for three reasons. First, the two-minute Just Drive task was approximately equal in duration to the auditory-vocal tasks (except for the Book-on-Tape Summarize task). Second, any distraction effects during Just Drive would likely be, like auditory-vocal tasks, of a primarily cognitive nature rather than due to other resource competition. Finally, Just Drive increased the task set size for auditory-vocal correlation and paired-comparison analyses.

Predictions of the auditory-vocal tasks were based in part on the previous model and also on a task content analysis.

Unlike visual-manual tasks, auditory-vocal tasks do not impose visual or manual demands likely to conflict with driving. Short-term or working memory processes (Nairne, 2003) would likely be the principal locus of task demands. MRT theory and modeling predicts that, relative to other auditory-vocal study tasks, Travel Computations and Route Orientation should be among the higher-workload auditory-vocal tasks. This is indicated because of their potential interference with driving through spatial working memory demands. For example, mental arithmetic such as that used in Travel Computations has been associated with spatial cognition (Baddeley, 1990).

The Route Orientation task was motivated by the research of Brooks (1968) that showed a spatial (mental imagery) task resulted in less conflict with a verbal response than a spatial response. Of the auditory-vocal tasks, Route Instructions stands out as another higher-workload task. The number of information items whose details needed to be recalled and recited, though verbal, were about left- or right turns (spatial content) and were difficult to remember. This task resembled a running memory task (Baddeley, 1990). The other auditory-vocal tasks were put into the lower-workload category because of their primarily verbal working memory load and their routine nature. Book-on-Tape Summarize was placed in the lower category because of the limited effort expected given modest levels of oral expression abilities (Carroll, 1993).

The remaining two tasks were mixed-modality tasks that include both visual-manual and auditory-vocal components. These were termed exploratory in

Figure 2-2 because the properties of mixed-mode tasks that involve interaction with an interactive voice response system are just beginning to be understood (e.g., Balentine and Morgan, 1999).

CAMP analytical modeling showed an interesting reversal of application. The DTCP values based on MRT showed much better prediction for the auditory-vocal task set than they did for the visual-manual task set. That is, the DTCP values accorded better, though not perfectly, with prior research in support of MRT. On the other hand, the CAMP analytic model activity times did not provide sensible output for the fixed-duration auditory-vocal and Just Drive tasks. For example, Sports Broadcast and Biographical Q&A had longer estimated activity times than Travel Computations or Route Orientation. Table 2-4 presents the basis of the auditory-vocal and Just-Drive task categorization.

Table 2-4. Basis of Relative Higher Versus Lower Workload Prior Prediction for Auditory-Vocal and Just Drive Tasks

Task	Prior Research	MRT Modeling – Median of Raters	Engineering Judgment	Final Prediction
Just Drive	---	---	L	Lower
Sports Broadcast	L	L	L	Lower
Biographical Q&A	L	H	L	Lower
Book-on-Tape Listen	L	L	L	Lower
Book-on-Tape Summarize	L	L	L	Lower
Travel Computations	H	H	H	Higher
Route Instructions	H	H	H	Higher
Route Orientation	H	H	H	Higher

2.4 Research Hypotheses and Their Validity

Stripped to the basics, drivers steer, operate accelerator and brake pedals, look at the road scene directly and through mirrors, detect objects and events so as to respond to them, and occasionally check gages. This is the basis of the CAMP DWM approach to measure task-induced driver

distraction. Assessments that flow from such an approach naturally point to measures of lateral control performance, longitudinal control performance, object-and-event detection, and eyeglance behavior. Prior research and theory suggest that different measurement categories can assess different aspects of task-related distraction.

Table 2-5 presents a summary of selected driving performance and eyeglance measures as well as the research hypotheses and rationale behind them. The table is not exhaustive of the many research questions that have, in fact, been pursued in this project. However, the table provides a sense of how the research hypotheses might be posed. In the examples given in the table, higher-workload tasks would be expected to be associated with higher values of the measures indicated. Chapter 1 contained additional examples of measures for which higher workload tasks would be associated with lower values of the measures relative to lower-workload tasks.

Table 2-6 presents a similar summary of selected laboratory surrogate measures as well as the research hypotheses and rationale behind them. These are a subset of measures that might be gleaned from the surrogate methods studied in this project. They are among those thought to be predictive of distraction while driving. The rationale provides some indication of the links that motivate interest in the associated surrogate measures.

Table 2-5. Driving Performance Measures and Eyeglance Behavior: Example Research and Rationale

Objective	Measure	Research Hypothesis	Rationale
Assess the relationship between In-Vehicle Tasks and Lateral Control	Standard Deviation of Lane Position during task (SDLP)	Higher workload > Lower workload	SDLP is a continuous, ever-present measure of lanekeeping that also shows differences among tasks. Normal probability theory suggests that larger SDLP implies an increased likelihood of departing the travel lane. While there may be no lane exceedances for a given trial, there is always lanekeeping to measure.
	Percent of participants with one or more lane exceeds during task %Lanex (Cross)	Higher workload > Lower workload	Lane Exceedances are discrete, infrequent events that provide an indication of egregious lapses in lanekeeping that differentiates tasks
Assess the relationship between In-Vehicle Tasks and Longitudinal Control	Difference between Minimum Speed and Maximum Speed for duration of Task (SpeedDiff)	Higher workload > Lower workload	Speed reduction is commonly associated with increased distraction in an attempt to increase the safety margin or reduce the demand of concurrent driving. The lead vehicle traveled at constant velocity (in cruise control). This made Speed a good proxy for time-headway and range. It was preferred in that it provided a more robust/reliable signal.
Assess the relationship between In-Vehicle Tasks and Object-	Lead Vehicle Deceleration (LVD) Miss Rate and Response Latency	Higher workload > Lower workload	Looming is an important visual stimulus to detect. Here is it exercised by lead vehicle coast-down deceleration (no brake lights).

Objective	Measure	Research Hypothesis	Rationale
and Event Detection (OED)	Center High-Mounted Stoplight Miss Rate and Response Latency	Higher workload > Lower workload	Another key visual stimulus in driving is light onset, e.g., traffic or brake light onset. Here, it is exercised by CHMSL onset in the lead car ahead.
	Follow-Vehicle Turn Signal (FVTS) Miss Rate and Response Latency	Higher workload > Lower workload	Situation awareness depends, in part, on mirror sampling. The FVTS onset provided a stimulus to look for in mirrors.
Assess the relationship between In-Vehicle Tasks and Selected Eyeglance Measures	Mean Total Glance Time to Task Related Areas	Higher workload > Lower workload	This measure reflects the overall visual demand of a task and the overall duration for which the driver was looking away from the driving scene.
	Mean Number of Glances to Task Related Areas	Higher workload > Lower workload	This measure is thought to reflect the complexity of a task as a whole, i.e., the number of task components
	Mean Duration of Glances to Task Related Areas	Higher workload > Lower workload	This measure is thought to indicate the difficulty of a visual task component. It may trade off with Number of Glances. Mean Glance Duration multiplied by Mean Number of Glances approximates Mean Total Glance Time.

Table 2-6. Laboratory and Surrogates: Example Research Hypotheses and Rational

Objective	Measure	Research Hypothesis	Rationale
Assess the Subjective Impressions that Tasks make on Test Participants	Operator Workload (OWL)	Higher workload > Lower workload	Subjective assessments of relative task demand may be related to actual performance and behavior.
	Multitasking Difficulty Magnitude Estimate	Higher workload > Lower workload	
	Situational Awareness Magnitude Estimate	Higher workload > Lower workload	
Assess Surrogates of Task Duration while concurrently driving	Static Time	Higher workload > Lower workload	Workload is often time-driven. Various measures of workload are themselves duration dependent.
	STISIM Task Duration	Higher workload > Lower workload	
Assess Surrogates of Visual Demand	Total Shutter Open Time (TSOT)	Higher workload > Lower workload	Higher TSOT is intended to reflect the visual demand of the task. It may reflect Task-related Glance Counts, EORT*, or MSGT**.
	Resumeability Metric (R-Metric)	Higher workload > Lower workload	Higher R-Metric values are thought to reflect difficulty in resuming visual tasks because of increased need to reorient to the task

Objective	Measure	Research Hypothesis	Rationale
Assess Surrogates of Object-and-Event Detection Performance	PDT-Alone MissRate and RT	Higher workload > Lower workload	Simple Event Detection (light onset) may be sufficiently predictive
	PDT-with-STISIM MissRate and RT	Higher workload > Lower workload	Simple Event Detection, but now under concurrent driving and task load
	Sternberg task (Various Miss, Error, and RT Measures)	Higher workload > Lower workload	Detection with a response choice (from memory set: yes/no). An attempt to tap into simple decision making
Assess Surrogates of Driving	STISIM Scenario	Various measures of Task Duration while driving, Lateral control, Longitudinal control	Part Task Simulator with a proven research record thought to provide useful proxies for related measures on road or track.

* EORT stands for Eye-Off-Road Time-Task-Related

** MSGT stands for Mean Single-Glance Time

A great many questions may be raised regarding these measures, the hypotheses attached to them, and the rationale proposed. Some of these questions are addressed in Chapter 8, *Discussion and Recommended Toolkit*.

2.5 Data Analysis

The DWM project was complicated to plan and execute. It addressed a complex and as yet poorly understood phenomenon called driver workload. Because of this, the DWM principal investigators sometimes took different perspectives with respect to statistical procedures, data partitioning, and treatment of results. This provided an opportunity to examine the robustness of findings. It also led to a wider range of interpretations and ways the data might be related to driver distraction. Different principal investigators examined different parts of the overall database from this study. Thus, different sections of this report will be accompanied by descriptions and rationale for the methods used for each section. In the end, what is common in the data generally outweighed what is different in the analysis approaches.

2.6 Chapter References

Balentine, B., and Morgan, D. P. (1999*). How to build a speed recognition application: A style guide for telephony dialogues*. San Ramon, CA: Enterprise Integration Group.

Baddeley, A. (1990). *Your memory: A user's guide*. London

Brooks, L.R. (1968). Spatial and verbal components in the act of recall. *Canadian Journal of Psychology*, 22, 349-368.

Card, S. K., Moran, T. P., and Newell, A. (1983). *The psychology of human-computer interaction*. Mahwah, NJ: Lawrence Erlbaum .

Carroll, J. B. (1993). *Human cognitive abilities: A survey of facto-analytic studies*. London: Cambridge University Press.

Della Sala S., Baddeley A., Papagno C., Spinnler H. (1995). Dual-task paradigm: a means to examine the central executive. *Annals of the New York Academy of Sciences*; 769, 161–71.

Goodman, M.J., Barker, J.A., and Monk, C.A. (February, 2005). *A bibliography of research related to the use of wireless telecommunications devices from vehicles*. Washington, DC: National Highway Traffic Safety Administration.

Goodman, M.J., Tijerina, L., Bents, F.D., and Wierwille, W.W. (1999). Using cellular telephones in vehicles: Safe of Unsafe *Transportation Human Factors Journal*, 1(1), 3-42.

Groeger, J. (2000). *Understanding driving: Applying cognitive psychology to a complex everyday task*. London: Taylor and Francis.

Harris, R. M., Iavecchia, H. P., Bittner, A. C., Jr. (1988). Everything you always wanted to know about HOS micromodels but were afraid to ask. *Proceedings of the Human Factors and Ergonomics Society 32nd Annual Meeting*, 1051-1055.

John, B. E., and Gray, W. D. (2000). *GOMS Tutorial/Workshop*. Presented at the 2000 IEA/HFES Conference. Available at http://hfac.gmu.edu/~gray/pubs/papers/goms/goms-wdg_dab.htm.

Laberge-Nadeau, C., Maag, U., Bellavance, F., Lapierre, S. D., Desjardins, D., Messier, S., and Saidi, A. (2003). Wireless telephones and the risk of road crashes. *Accident Analysis and Prevention*, 35, 649-660.

Laughery, K. R., Jr., Archer, S., and Corker, K. (2001). Modeling human performance in complex systems. In G. Salvendy (Ed.), *Handbook of industrial engineering, Third edition* (pp. 2409-2444).

Lewis, J. R. (1988). Pairs of Latin squares to counterbalance sequential effects and pairing of conditions and stimuli. *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1223-1227). Santa Monica, CA: Human Factors and Ergonomics Society.

Niebel, B. W. (1976). *Motion and time study (Sixth Edition)*. Homewood, IL: Richard D. Irwin, Inc.

Nairne, J. S. (2003). Sensory and working memory. In A. F. Healy and R. W. Proctor (Eds.), *Handbook of psychology: Volume 4-Experimental psychology* (423-444). New York: John Wiley.

Nowakowski, C., and Green, P. (2001). *Prediction of menu selection times parked and while driving using the SAE J2365 Method* (Technical Report 2000-49). Ann Arbor, MI: University of Michigan Transportation Research Institute.

Redelmeier, D. and Tibshirani, R. (1997). Association between cellular telephone calls and motor vehicle collisions. *The New England Journal of Medicine*, *336*(2), 453-458.

Redelmeier, D. and Tibshirani, R. (2003). Is using a cell phone like driving drunk? In J. P. Rothe (Ed.), *Driving lessons: Exploring systems that make traffic safer*. Edmonton, Alberta: University of Alberta Press.

Smith, G. L., Jr. (1978). *Work measurement: A systems approach*. Columbus, OH: Grid Publishing.

Stutts, J., Reinfurt, D.W., Staplin, L, and Rodgman, E.A. (2001). *The role of driver distraction in traffic crashes*. New York: AAA Foundation for Traffic Safety.

Tijerina, L., Parmer, E., and Goodman, M. J. (1999). Individual differences and in-vehicle distraction while driving: A test track study and psychometric evaluation. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*. 982-986.

Violanti, J. M. and Marshall, J. R. (1996). Cellular phones and traffic accidents: an epidemiological approach. *Accident Analysis and Prevention*, *28*, 265-270.

Wang, J-S., Knipling, R.R., and Goodman, M. J. (1996). The role of driver inattention in crashes: New statistics from the 1995 Crashworthiness Data System. *The 40th Annual Proceedings: Association for the Advancement of Automotive Medicine*, 377-392.

Wickens, C. D., and Hollands, J. G. (2000). *Engineering psychology and human performance* (Third edition). Upper Saddle River, NJ: Prentice Hall.

3 Test Track Results

3.1 Background

Chapter 2, *Study Design Overview*, described the procedures and tasks used in the CAMP DWM Project during test track testing at Ford Motor Company's Michigan Proving Ground in Romeo, Michigan. Details of the materials, equipment, and procedures used are provided in appendices to this report.

This chapter presents the results of the test track work. Included are the task effects on object and event detection, driver eyeglance behavior, and lateral and longitudinal vehicle control. Summary statistics for all measures reported in this chapter are provided in the appendices.

3.2 Test Track Participants

An independent marketing firm recruited 64 licensed drivers from the Detroit metropolitan area for participation in the test track phase of the study. The participants spanned six age ranges: 21 to 29, 30 to 39, 40 to 49, 50 to 59, 60 to 69, and 70 to 79 years old.

The prospective candidates were screened for good health and a good driving record before being admitted into the study. Thirty-three of the participants were female and 31 were male. Table 3-1 presents the distribution of the age and gender of the participants. The participants were paid $400 for their two-day time commitment to the study. Additional details about the sample of subjects and the screening process are provided in the appendices.

Table 3-1. Age and Gender of Test Track Participants

	Age Category						
	20's	30's	40's	50's	60's	70's	All
Male	6	4	5	6	5	5	31
Female	5	7	7	5	6	3	33
All	11	11	12	11	11	8	64

3.3 Test Track Task Effects on Object-and-Event Detection

As part of test-track and on-road driving, the participants were presented with three types of object-and-event detection (OED) roadway events.

The first type of OED stimulus was the Center High-Mounted Stoplight event. The CHMSL of the lead vehicle would illuminate for a duration equal to the time headway between the subject vehicle and the lead vehicle. The lead vehicle, however, did not decelerate during the CHMSL event in order to roughly simulate "riding the brakes." The participants responded if they detected the CHMSL event. The reaction time to detect the CHMSL event was recorded and the percentage of missed detections was calculated.

The second type of OED stimulus was the Lead Vehicle Deceleration (LVD) event. From time to time during test trials, the lead vehicle would begin to slow. However, the brake lights of the lead vehicle would not illuminate so as to simulate a "coast down" maneuver. If the participants detected this, they were to gently tap the brake pedal as soon as they could. The lead vehicle resumed its speed to 55 mph and the trial continued. The participants were instructed to accelerate, if necessary, at the conclusion of a trial to close the gap between the subject vehicle and the lead vehicle after this event was completed. The reaction time to detect the LVD event was recorded and the percentage of missed detections was calculated.

The third type of OED stimulus was the Follow Vehicle Turn Signal (FVTS) event. Periodically, the follow vehicle would illuminate its driver-side front (left-front) turn signal for 2.5 seconds to simulate a follow vehicle's indication to overtake. FVTS events were primarily seen in the left outside mirror and, occasionally, in the inside rearview mirror. The participants were asked to detect the FVTS event and respond if the signal was detected. The reaction time to detect the FVTS event was recorded and the percentage of missed detections was calculated.

Response time (RT) was calculated from the stimulus onset time to the participant's response. There was a 200-mseconds lockout after stimulus onset to prevent anticipation responses, and the participant then had two seconds to respond. The participant was asked to respond to the OED stimulus during the duration of the subsidiary task. If the participant responded to the OED event after the task was completed, that response was not recorded and the trial was recorded as a missed detection. CHMSL and FVTS stimulus onsets began when a computer signal from the data acquisition system reached the follow or lead vehicle, as appropriate. LVD stimulus onset was from receipt of a signal to the lead vehicle to disable the cruise control and begin the coast down if driving conditions permitted it.

Participants were scheduled to perform each DWM task with each of the three OED events (e.g., CHMSL, FVTS, or LVD). Each test participant was scheduled to perform a task twice with a given OED stimulus on the track. Each OED presented was scored either "detected" or "not detected" based on the operational definitions used in the study. Detections were coded as one (1). Missed Detections were coded as zero (0). The binary (0, (1) detection data from the two trials for a participant were then averaged. This was done for each OED stimulus for each task. If there were no detections, the resulting participant score was 0.0. If only one trial was completed and a detection occurred, the participant score was 1.0 and 0.0 otherwise. If two trials were completed, the participant detection score was 0.5 if there was a detection on one trial and a missed detection on the other trial. If there were detections on both trials, the participant detection score was 1.0. The average of these per-participant detection scores, averaged across all participants, was taken as the proportion of "Detections." The Proportion of Detections was multiplied by 100 to create "Percent Detections." The "Percent Missed Detections" variable was defined as 100 percent minus the Percent Detections. This approach was motivated by the fact that the arithmetic mean (average) of a sample of binary (1, 0) scores is the proportion of scores coded as 1.

3.3.1 Center High-Mount Stoplight (CHMSL) Results

The results from the test-track CHMSL events are given in Figure 3-1. Overall, Percent Missed Detections were between 10 percent and 35 percent. Visual-manual tasks generally had higher missed detection rates (between 25% and 35%) than the auditory-vocal and Just Drive tasks (between 15% and 25%, approximately). Some auditory-vocal and mixed-mode tasks (Route Instructions, Sports Broadcast, and Delta Flight Information (Delta Flightline), respectively) had even slightly lower miss rates than Just Drive alone. Higher CHMSL miss rates for visual-manual tasks were likely due to the need to look away from the road scene while completing the task. Participants looked down or otherwise away from the road scene to execute the visual-manual tasks. Thus, foveal vision to the roadway would momentarily be removed and leave only peripheral vision to detect CHMSL onset. Task duration also appeared to play a role in object and event detection. Shorter visual-manual and auditory-vocal tasks (e.g., Book-on-Tape Summarize) showed poorer detection performance than the longer auditory-vocal, Just Drive, and mixed-mode (e.g., Delta Flight Information) tasks. Furthermore, detection was not perfect in any task condition, not even Just Drive. In fact, some auditory-vocal tasks actually had lower CHMSL miss rates than Just Drive. This suggests that driver distraction of the "lost in thought" variety may have been reduced by the presence of an in-vehicle activity.

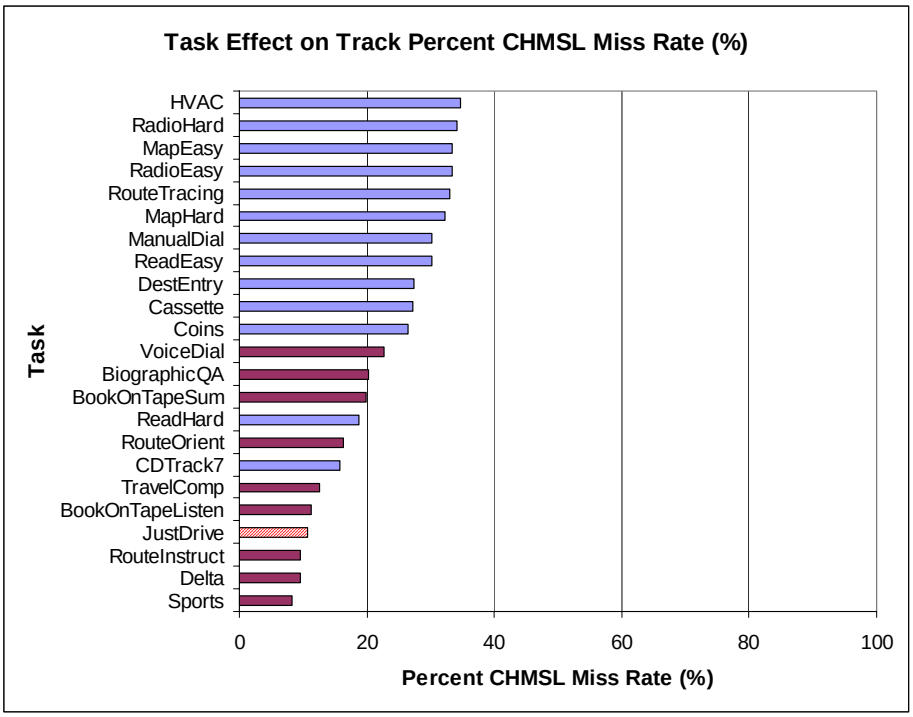


Figure 3-1. Track Percent CHMSL not Detected (Missed)

3.3.2 Lead Vehicle Deceleration (LVD) Results

The results from track testing for the LVD event showed a similar pattern to that of the CHMSL event (see Figure 3-2). Visual-manual tasks generally showed a higher missed detection rate than most auditory-vocal tasks. However, two visual-manual tasks (Read (Hard) and CD/Track (7) had LVD miss rates that were interspersed among the auditory-vocal and Just Drive tasks. The Just Drive task had nearly the lowest percentage of missed detections, although three auditory-vocal tasks (the same as those found with CHMSL detection) had slightly lower missed detections than Just Drive.

It should be noted that a portion of the LVD events were not detectable within the task length of short visual-manual tasks. One possible explanation for this may be that in short duration tasks the stimulus is below the optical looming threshold for detection. This issue is discussed further in Chapter 8, *Discussion and Summary*.

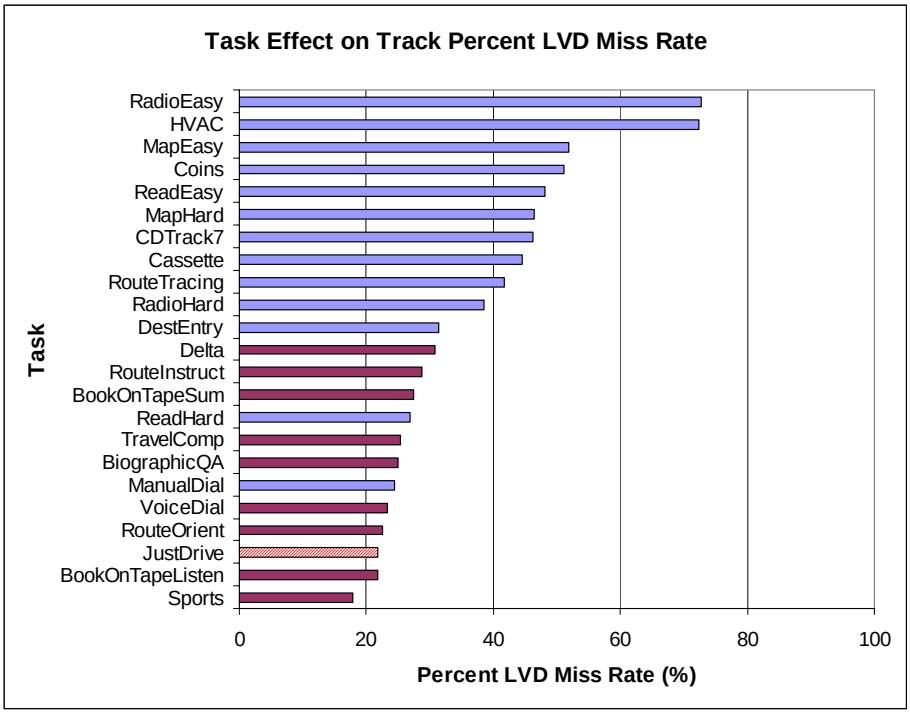


Figure 3-2. Track Percent LVD Not Detected (Missed)

3.3.3 Follow Vehicle Turn Signal (FVTS) Results

The results from the FVTS event are given in Figure 3-3. FVTS miss rates were in the 50 percent to 90 percent range, which was higher than those of the CHMSL or LVD events, overall. The visual-manual tasks generally had higher missed detection rates than auditory-vocal tasks. However, Read (Hard) and Manual Dial - Home miss rates were interspersed among the auditory-vocal and Just Drive tasks (see Figure 3-3).

Some observers felt that the FVTS event placed an unrealistic emphasis on events that occurred to the rear of the subject vehicle. The concern was that the FVTS stimulus would focus much more attention by the driver to the inside and outside rearview mirrors than is likely to occur in real-world driving. The high missed-detection rates for the FVTS event indicated that a hyper-focus on the rearview mirrors did not occur. During secondary task loading, drivers appeared to have prioritized the forward road scene over the rearward visual scene.

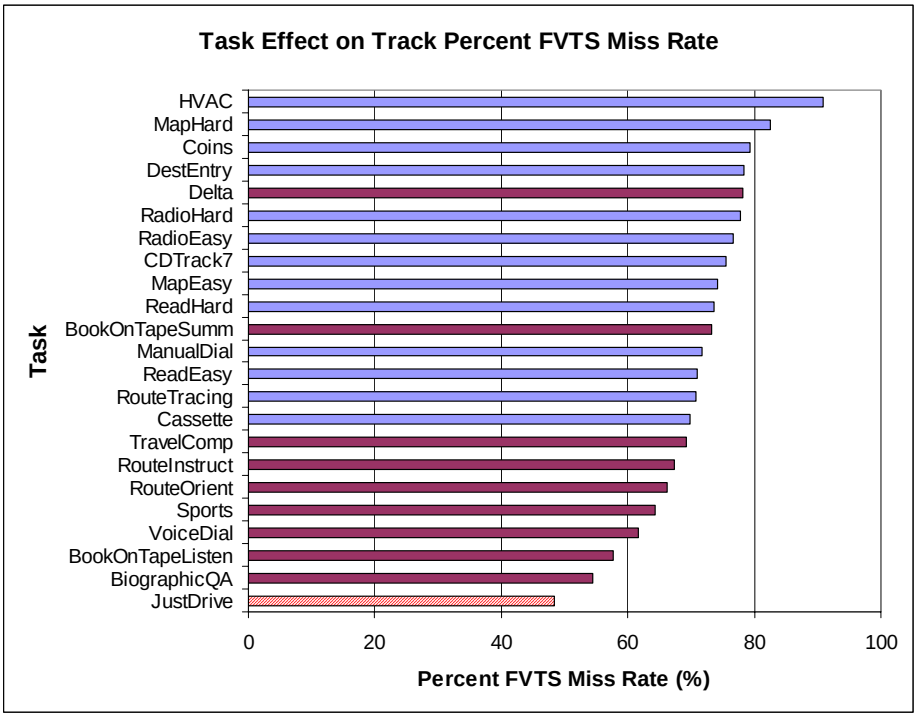


Figure 3-3. Track Percent FVTS Not Detected (Missed)

3.4 Test Track Task Effects on Glance Behavior

Analyses of driver eyeglance behavior examined multiple metrics related to the locations, durations, and rates of glances during the performance of tasks by drivers in this study.

A detailed description of the methods for obtaining and reducing data on eye movements is provided in Appendix P. Briefly, driver eye-movement data were derived from the digital-video images recorded at 30 frames per second through the camera view of the drivers' faces during data collection. The locations and durations of glances were determined by human analysts scoring the video with The Observer software from Noldus Information Technology. For purposes of the analyses reported, data were extracted from videos for 18 test track participants. The participant sample was balanced by gender (9 males, 9 females) and by age group. Within each gender, there were three individuals in each of three age categories: younger (20 to 39), middle-age (40 to 59), and older (60 to 79). The sample of participants for eyeglance data reduction was selected for the quality of eyeglance video available within the age and gender categories.

For each video scored, the location of each glance was determined, one of nine zones was assigned (see Appendix P), and the duration was measured. This was done independently by two analysts, to assure accurate scoring. A mediator compared the location and duration measurements of the two analysts, and then resolved any discrepancies of location or timing (equal to or greater than 3 frames) by creating a final composite scoring of each data file. The final data files were exported from The Observer software, merged with the driving performance data, and imported into SAS for statistical analysis.

For analyses of task effects, the original nine locations of glances were mapped to the following four location types:

- **Road** glances to the roadway location (in this chapter the roadway location refers to the test track)
- **Situation Awareness (SA)** glances to any mirror and to the speedometer (for other situation awareness glances)
- **Task** glances to all task-related locations
- **NA** glances that were missing, obstructed (e.g., driver held a map in front of his/her eyes), or otherwise not able to be scored or attributable to one of the above categories. This is essentially residual task time with unknown eyeglance behavior.

For special analyses, glance locations associated with mirrors (outside left, outside right, and inside rearview mirror) were also mapped to a mirror location type. For other analyses, the three categories of task, SA, and NA were grouped into a Not Road (NR) category. The inclusion of the NA category is based on the assumption that NA glances were not to the road.

Multiple metrics were examined for glances to each type of location. These included metrics based on number of glances, duration of glances, rate of glances per second, proportion of task duration spent looking at a location, and accumulated durations (such as total time at each location). The full list of eyeglance metrics and their associated definitions are shown in Table 3-2.

Table 3-2. Eyeglance Metrics: Name and Definition of Metrics Used in Study

	Eyeglance Metric / Definition
1	**MeanTskglncs_ allc:** Mean Number of Task Glances. Total number of glances that occurred during the task.
2	**MeanTaskdur_ allc:** Mean Task Duration. Mean task duration computed from eyeglance data.
3	**MeanmeanTdur_ allc:** Mean Glance Duration. Mean of the mean durations of glances of all types during task.
4	**MeanmedTdur_ allc:** Mean Median Glance Duration. Mean of median durations of glances of all types during task.
5	**MeansdTdur_ allc:** Mean Standard Deviation of Glance Duration. Mean of the standard deviations of duration of glances of all types during task.

<table>
<tr><th></th><th>Eyeglance Metric / Definition</th></tr>
<tr><td>6</td><td>MeanTglsprs_ allc: Mean Glance per Second. Mean rate of glances of all types per second during task.

Eye glances to the nine areas in the car and the missing classification when combined with the Task/Non-Task modifier yield 20 locations numbered 0 to 19. These glances are then reduced into 6 classes of glances. The locations are included in each of the collapsed groups, as shown:
<table>
<tr><th>Class of Glances</th><th>Individual Locations</th><th>Task/Non-Task</th></tr>
<tr><td>Road or RD</td><td>Forward Road Scene</td><td>Non-Task</td></tr>
<tr><td>Situational Awareness or SA</td><td>Steering Wheel/IP, Left Mirror, Right Mirror, Rearview Mirror</td><td>Non-Task</td></tr>
<tr><td rowspan="2">Task Related or TR</td><td>Forward, Steering Wheel/IP, Down, Center Console, Up Visor</td><td>Task</td></tr>
<tr><td></td><td></td></tr>
<tr><td rowspan="2">Not Applicable or NA</td><td>Down, Missing, Center Console, Up Visor, Other</td><td>Non-Task</td></tr>
<tr><td>Missing, Left Mirror, Right Mirror, Rearview Mirror, Other</td><td>Task</td></tr>
<tr><td rowspan="2">Mirror Related or MR</td><td>Left Mirror, Right Mirror, Rearview Mirror</td><td>Non-Task</td></tr>
<tr><td>Left Mirror, Right Mirror, Rearview Mirror</td><td>Task</td></tr>
<tr><td rowspan="2">Not Road or NR</td><td>Steering Wheel/IP, Down, Missing, Left Mirror, Center Console, Right Mirror, Up Visor, Rearview Mirror, Other,</td><td>Non-Task</td></tr>
<tr><td>Forward Road Scene, Steering Wheel/IP, Down, Missing, Left Mirror, Center Console, Right Mirror, Up Visor, Rearview Mirror, Other,</td><td>Task</td></tr>
</table></td></tr>
<tr><td>7</td><td>MeangIncesRD_ allc: Mean number of glances to the forward road scene during task.</td></tr>
<tr><td>8</td><td>MeanduratRD_ allc: Mean duration of all glances to the forward road scene during task (summed across all glances).</td></tr>
<tr><td>9</td><td>MeanmeanRDdur_ allc: Mean of the mean duration of glances to the forward road scene during task.</td></tr>
<tr><td>10</td><td>MeanmedRDdur_ allc: Mean median duration of glances to the forward road scene during task.</td></tr>
<tr><td>11</td><td>MeansdRDdur_ allc: Mean standard deviation duration of glances to the forward road scene during task.</td></tr>
<tr><td>12</td><td>MeangrateRD_ allc: Mean rate of glances to the forward road scene per second during task.</td></tr>
<tr><td>13</td><td>MeanpctdurRD_ allc: Mean proportion of time glancing to the forward road scene during task.</td></tr>
<tr><td>14</td><td>MeangIncesSA_ allc: Mean number of glances to situational awareness locations during task.</td></tr>
<tr><td>15</td><td>MeanduratSA_ allc: Mean duration of all glances to situational awareness locations during task (summed across all glances).</td></tr>
<tr><td>16</td><td>MeanmeanSAdur_ allc: Mean of the mean duration of glances to situational awareness locations during task.</td></tr>
</table>

	Eyeglance Metric / Definition
17	**MeanmedSAdur_ allc:** Mean median duration of glances to situational awareness locations during task.
18	**MeansdSAdur_ allc:** Mean standard deviation duration of glances to situational awareness locations during task.
19	**MeangrateSA_ allc:** Mean rate of glances to situational awareness locations per second during task.
20	**MeanpctdurSA_ allc:** Mean proportion of time glancing to situational awareness locations during task.
21	**MeanglncesTR_ allc:** Mean number of glances to task-related locations during task.
22	**MeanduratTR_ allc:** Mean duration of all glances to task-related locations during task (summed across all glances).
23	**MeanmeanTRdur_ allc:** Mean of the mean duration of glances to task-related locations during task.
24	**MeanmedTRdur_ allc:** Mean median duration of glances to task-related locations during task.
25	**MeansdTRdur_ allc:** Mean standard deviation duration of glances to task-related locations during task.
26	**MeangrateTR_ allc:** Mean rate of glances to task-related locations per second during task.
27	**MeanpctdurTR_ allc:** Mean proportion of time glancing to task-related locations during task.
28	**MeanglncesNA_ allc:** Mean number of glances to not applicable locations during task.
29	**MeanduratNA_ allc:** Mean duration of all glances to not applicable locations during task (summed across all glances).
30	**MeanmeanNAdur_ allc:** Mean of the mean duration of glances to not applicable locations during task.
31	**MeanmedNAdur_ allc:** Mean median duration of glances to not applicable locations during task.
32	**MeansdNAdur_ allc:** Mean standard deviation duration of glances to not applicable locations during task.
33	**MeangrateNA_ allc:** Mean rate of glances to not applicable locations per second during task.
34	**MeanpctdurNA_ allc:** Mean proportion of time glancing to not applicable locations during task.
35	**MeanglncesMR_ allc:** Mean number of glances to mirror-related locations during task.
36	**MeanduratMR_ allc:** Mean duration of all glances to mirror-related locations during task (summed across all glances).
37	**MeanmeanMRdur_ allc:** Mean of the mean duration of glances to mirror-related locations during task.

	Eyeglance Metric / Definition
38	**MeanmedMRdur_ allc:** Mean median duration of glances to mirror-related locations during task.
39	**MeansdMRdur_ allc:** Mean standard deviation duration of glances to mirror-related locations during task.
40	**MeangrateMR_ allc:** Mean rate of glances to mirror-related locations per second during task.
41	**MeanpctdurMR_ allc:** Mean proportion of time glancing to mirror-related locations during task.
42	**MeanglncesNR_ allc:** Mean number of glances to all locations other than to the forward road scene during task.
43	**MeanduratNR_ allc:** Mean duration of all glances to locations other than to the forward road scene during task (summed across all glances).
44	**MeanmeanNRdur_ allc:** Mean of the mean duration of glances to locations other than to the forward road scene during task.
45	**MeanmedNRdur_ allc:** Mean median duration of glances to locations other than to the forward road scene during task.
46	**MeansdNRdur_ allc:** Mean standard deviation duration of glances to locations other than to the forward road scene during task.
47	**MeangrateNR_ allc:** Mean rate of glances to locations other than to the forward road scene per second during task.
48	**MeanpctdurNR_ allc:** Mean proportion of time glancing to locations other than to the forward road scene during task.
49	**MinTdur_ allc:** Minimum Glance Duration. Duration of shortest glance of all types during task.
50	**MinRDdur_ allc:** Duration of shortest glance to the forward road scene during task.
51	**MinSAdur_ allc:** Duration of shortest glance to situational awareness locations during task.
52	**MinTRdur_ allc:** Duration of shortest glance to task-related locations during task.
53	**MinNAdur_ allc:** Duration of shortest glance to not applicable locations during task.
54	**MinMRdur_ allc:** Duration of shortest glance to mirror-related locations during task.
55	**MinNRdur_ allc:** Duration of shortest glance to forward road scene during task.
56	**MaxTdur_ allc:** Maximum Glance Duration. Duration of Longest glance of all types during task.
57	**MaxRDdur_ allc:** Duration of Longest glance to the forward road scene during task.
58	**MaxSAdur_ allc:** Duration of Longest glance to situational awareness locations during task.
59	**MaxTRdur_ allc:** Duration of Longest glance to task-related locations during task.
60	**MaxNAdur_ allc:** Duration of Longest glance to not applicable locations during task.

	Eyeglance Metric / Definition
61	**MaxMRdur_ allc:** Duration of Longest glance to mirror-related locations during task.
62	**MaxNRdur_ allc:** Duration of Longest glance to forward road scene during task.
63	**N_EyeDataTasks_ allc:** This is a count that indicates the number of tasks with eyeglance data that contributed to the measures above.

3.4.1 Task Effects on Eyeglance Metrics

Analyses were undertaken on the data collected in the test track venue (separate from data collected in the on-road venue). Formal statistical analyses were conducted by, or with the assistance of, Carol Flannagan at the University of Michigan Traffic Research Institute. The data for each dependent measure were analyzed separately using linear mixed models (Verbeke and Molenberghs, 1997). This relatively new maximum-likelihood technique is ideally suited for an unbalanced design with multiple variance components. Subjects were treated as a random effect and the effects listed below were treated as fixed effects in the analysis.

- Between-subject factors treated as fixed effects were:
  - Age Group (young, middle, older)
  - Gender (male, female)
- Within-subject factors treated as fixed effects were:
  - Task (within type of task)
  - Location Type for Glances (road, situation awareness, task, NA)

Table 3-3 shows the results of the analyses performed across a variety of metrics. As is apparent, there was a significant main effect of Task on all of the glance metrics, as well as significant main effects of Location Type, and a significant Task by Location Type interaction.

Table 3-3. Linear Mixed-Model Effects for Glance Metrics

	Metrics								
Venue	**Effect**	**No. of Glances**	**Total Time At Location**	**Max Dur**	**Min Dur**	**Mean Dur**	**Med Dur**	**St Dev Dur**	**Glance Rate**
Test Track	Gender								
	Task	▲	▲	▲		▲	▲	▲	▲
	AgeGroup								
	Locat	▲	▲	▲	▲	▲	▲	▲	▲
	Task*AgeGroup	▲	▲						
	Task*Locat	▲	▲	▲		▲	▲	▲	▲
	AgeGroup*Locat	▲	▲						

Note: The red triangles indicate effects significant at $p < 0.05$.

Exploring the Task by Location Type interaction effects for these (and some other) variables is most meaningfully done graphically, starting with the metric *of Number of Glances*.

In the following sections, "Task" refers to up to 23 individual tasks evaluated on the test track. This is distinct from the Task Type factor analyzed with the mixed linear models. The "Task Group" factor had four levels: visual-manual, auditory-vocal, mixed-mode, and Just Drive. The data for all tasks within each level of task type were statistically combined for subsequent analyses. Individual DWM tasks were not analyzed in this way.

3.4.1.1 Number of Glances

Figure 3-4 shows the metric of Number of Glances by each Task and Location Type for the test track. It is apparent that there is a great deal of variation across tasks in Number of Glances per task. This might be expected because tasks were of different durations and longer task durations would allow more glances to be made. However, the variation across tasks also is a function of the type of location at which gaze is directed (road, situation awareness, or task-related). Within the interaction of Task by Location Type, there are several sub-patterns of interest. Focusing first on glances to the road, the region to which the highest number of glances were made, the Just Drive task led to more than 30 glances on average. A number of the auditory-vocal tasks (at the right side of the figure and highlighted with a dark red bar beneath the task names and numbers) produced nearly as many glances as Just Drive (between 20 and 35), with one exception—Book-on-Tape Summarize. This task was far shorter than all other auditory-vocal tasks (averaging about 20 seconds rather than 2 min) and was associated with fewer glances. For the Just Drive and the auditory-vocal tasks, the number of glances to the SA category parallels the number of glances to the road and the points lie virtually on top of the points for the number of glances to the road. Thus, the patterns of glancing to the road and SA location types for auditory-vocal tasks resemble the pattern for Just Drive.

For visual-manual tasks, the pattern is different. First, the number of glances to the road is dramatically lower for most of the visual-manual tasks—on-half to one-third of the number of glances to the road for Just Drive. This too is to be expected on the basis of task duration alone. Shorter tasks would allow fewer glances to be made overall. In addition for visual-manual tasks, glances to the SA category were fewer than to the road, less than half the number that were made to the road, on average, with the exception of Destination Entry, which differs from the pattern for auditory-vocal tasks. The Destination Entry task is an exception and, as the most visually intensive of the visual-manual tasks, required many more task-related glances to complete.

Therefore, drivers tended to make short glances, and to look frequently back to the road, thus increasing the number of glances to the road (over 50 on average). The number of task-related glances was, correspondingly also higher. Glances to mirrors and the speedometer (i.e., the SA category), were only slightly over 10 per task performance on average. For all other visual-manual tasks, the number of task-related glances was slightly less, but tended to closely track road glances. For visual-manual tasks, drivers usually glanced back and forth between the road and task.

Auditory-vocal tasks are not normally thought of as requiring any task-related glances, yet there were a few glances scored as task-related. These glances were noticed by the analysts who reduced the video eyeglance data. The glances that occurred tended to be glances up—either to the rear-view mirror in some of the language-production tasks, such as Biographical Q&A, or to the visor area. The glances that were to the rearview mirror area were clearly distinguishable from mirror glances made for the purpose of checking traffic insofar as they typically involved different body movement and head movement. They appeared to be glances at the rear-seat experimenter, as if the driver were seeking to make eye contact via the mirror with the person to whom they were speaking. However, no auditory-vocal task involved talking with the experimenter during a task's duration. Similarly, the looks to the visor area were distinct and recognizably different from other glances. One hypothesis for the looks to the visor area is that when a task requires the use of working memory, drivers look up, as if to visualize or “look at” the contents of working memory. Those auditory-vocal tasks for which task-related glances occurred were those that required retrievals from long-term memory, mental calculation, rehearsal in memory, or generation of linguistic material. There were no task-related materials in the mirror or visor areas. Nonetheless, it is possible that glances to these areas were associated with mental operations noted above or habits associated with speaking. Because of this, glances up during tasks were assumed to be task-related and were recorded and scored as such. These two types of glances were quite rare, but did occur for some types of auditory-vocal tasks.

In the figures that follow, data are shown on a particular glance metric for each of several glance locations: glances to the roadway (in this chapter, understood to be the test track), to the task, etc. The glance metric, such as Number of Glances, is shown on the y-axis, and individual tasks are shown on the x-axis. In these figures, colored lines connect data associated with glances to a particular location across different and discrete tasks. Lines, rather than bars, were used to facilitate understanding. It is much easier to follow an effect for glances to a particular location when the points are color-coded and connected by a line.

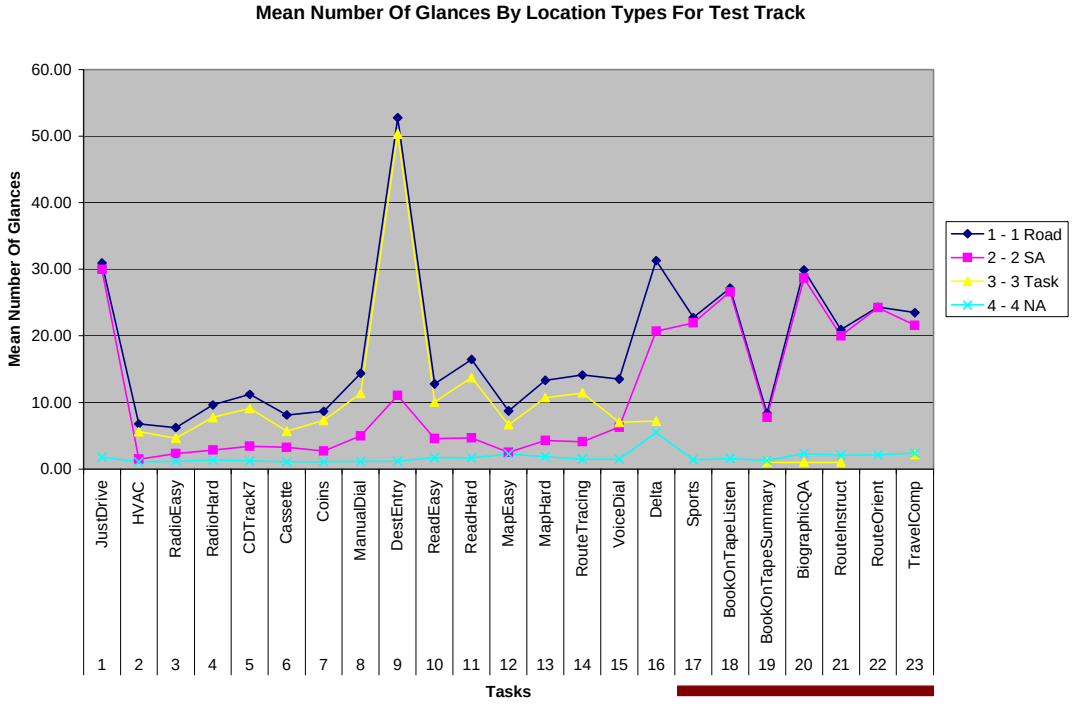


Figure 3-4. Test Track Mean Number of Glances by Task and Location Type

3.4.1.2 Glance Duration by Task and Location Type

Figure 3-5 shows Median Glance Durations, Figure 3-6 shows Mean Glance Durations, and Figure 3-7 shows Maximum Glance Durations. All three depict similar patterns. Glances to the road were much longer than glances to other regions for Just Drive and auditory-vocal tasks. For the Just Drive task, glances to the road tended to be 6 seconds (median) to 8 seconds (mean) in length. For auditory-vocal tasks, in which the eyes could be forward and on the road for the entire task, glances to the road tended to be longer (most of them in the range from about 7 to 14 seconds, based on medians) or (from 9 to 16 seconds, based on means). The extended length of these roadway glances raises a question about what prolonged gazes at the forward roadway may reflect about the focus of attention during these tasks. This issue is addressed later in this chapter. The typical lengths of glances for mixed-mode tasks fell in or just below the range of Just Drive. However, all visual-manual tasks had glance durations to the road and all other areas less than about 2 seconds in duration. Figure 3-8 shows mean glance durations for only task-related and situation-awareness glances, so that the scale of the figure could be enlarged within the region of these glance durations. With this scale change, it is possible to see that most task-related glances are between 0.8 and 1.40 seconds, on average, while most situation awareness glances averaged between 0.4 and 0.7 seconds in duration.

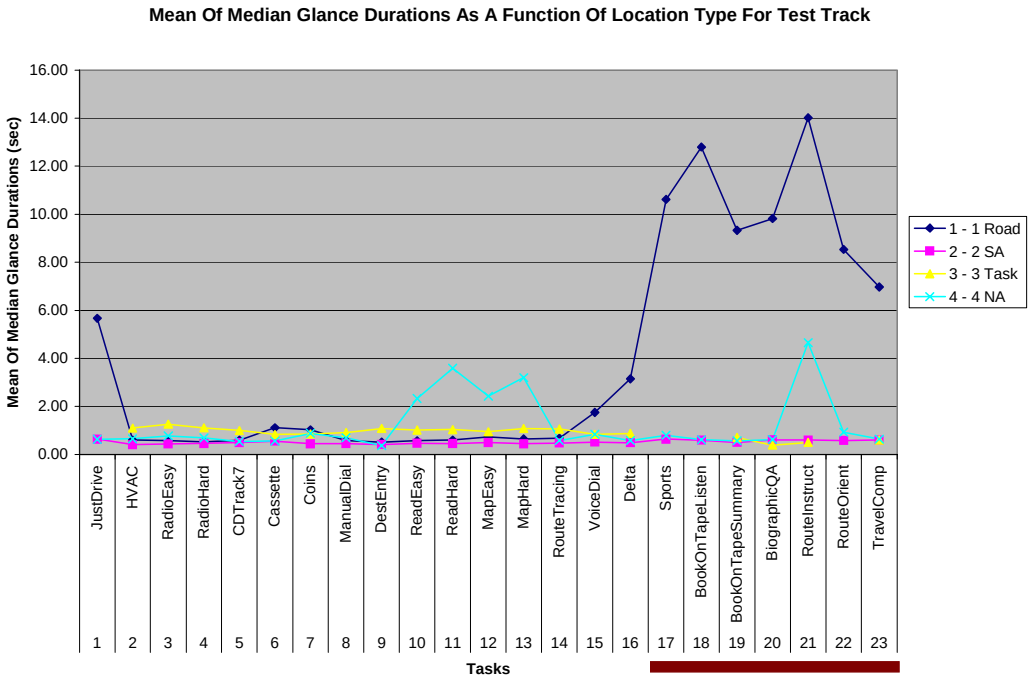


Figure 3-5. Test Track Mean of Median Glance Durations by Task and Location Type

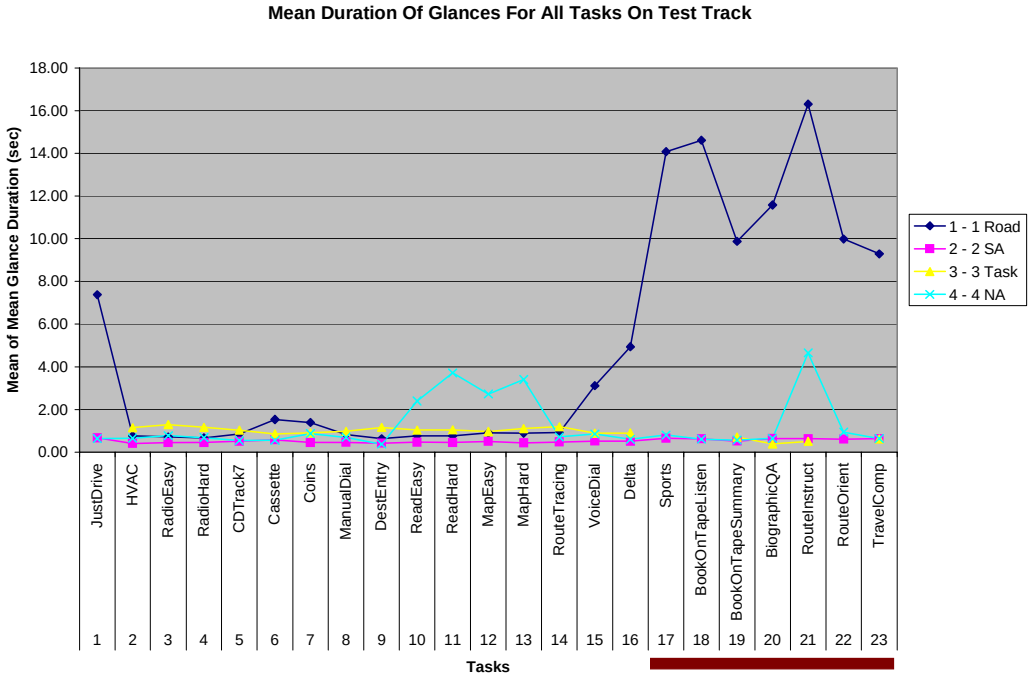


Figure 3-6. Test Track Mean of Mean Glance Durations by Task and Location Type

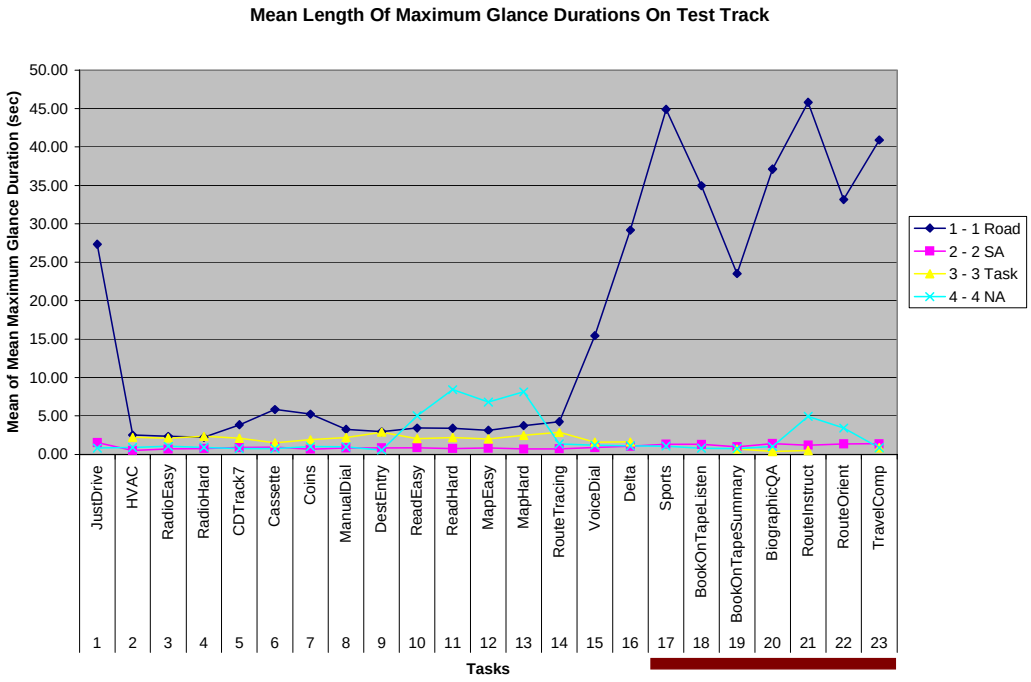


Figure 3-7. Test Track Mean of Mean Maximum Glance Durations by Task and Location Type

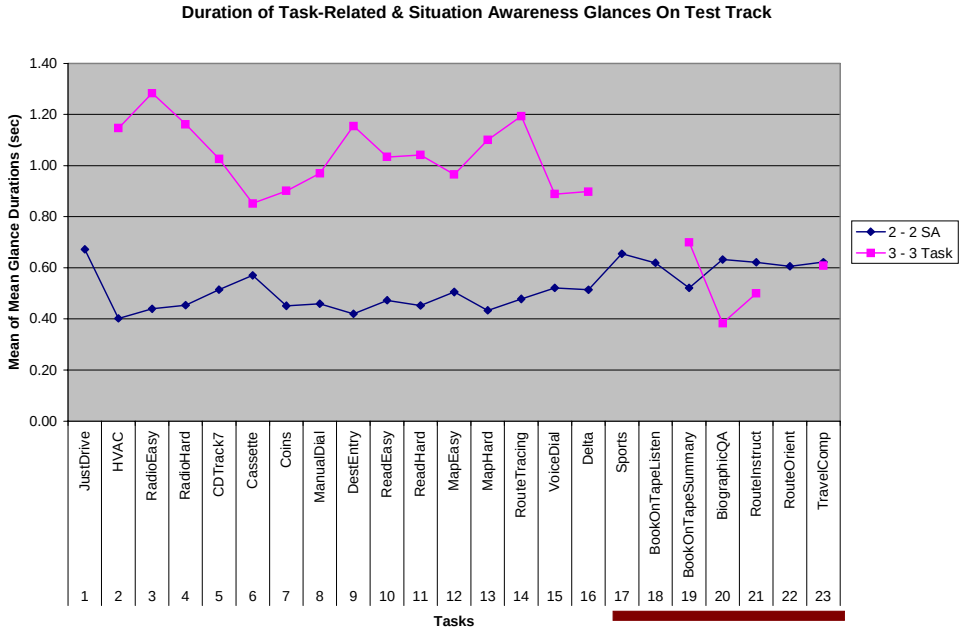


Figure 3-8. Test Track Mean of Mean Glance Duration of Task-Related and SA Glances by Task

Figure 3-9 shows Minimum Glance Durations by task and location type. As can be seen, the durations of the shortest glances to all locations were about 0.50 seconds in duration, with the exception of glances to the road for some of the auditory-vocal tasks (and for these, the minimums tended to be longer, approaching 2 seconds in duration).

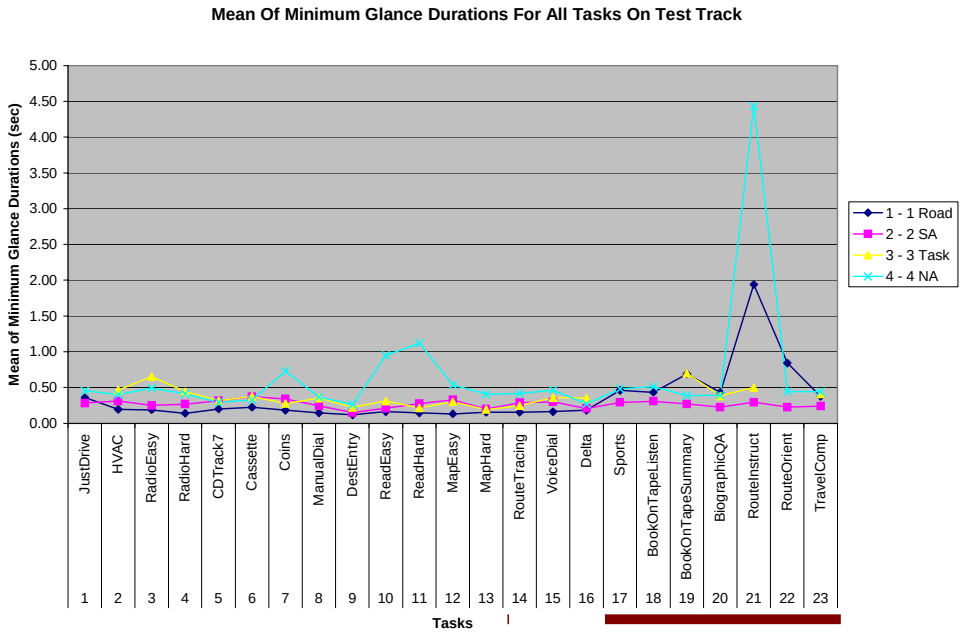


Figure 3-9. Effect of Location Type on Track Mean of Minimum Glance Durations by Task

3.4.1.3 Glance Rate by Task and Location Type

Figure 3-10 shows the interaction of Task by Location Type on the *Rate of Glancing to Each Location Type*. Visual-manual tasks produced the highest glance rates to the road (0.5 to 0.6 glances per seconds), as contrasted with 0.2 to 0.3 glances per seconds for Just Drive and the auditory-vocal tasks. Interestingly, the glance rate to mirrors and speedometer (SA is relatively stable across all tasks, at about 0.2) and the rate of glancing at the task were relatively stable for visual-manual tasks (between ~0.35 and ~0.52 glances per second, dropping off dramatically for mixed-mode tasks and auditory-vocal tasks).

Glance rates must be interpreted with caution. Simply because Task A had a higher glance rate to the road (e.g., 0.5 glances/per seconds) than Task B (e.g., 0.25 glances per seconds) does not necessarily mean that Task A was associated with higher levels of road monitoring than Task B. This is especially true for tasks with fewer but longer glances to the road. Glance rates must be considered in conjunction with the number of glances and duration of glances made during a task.

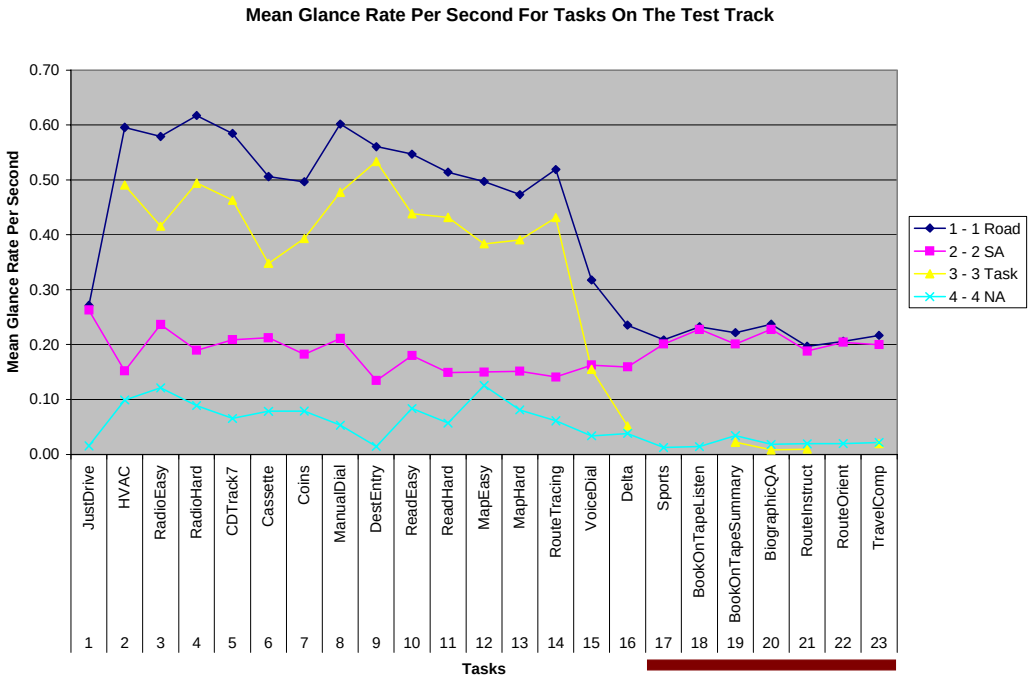


Figure 3-10. Test Track Rate of Glancing to Each Location Type by Task

3.4.1.4 Proportion of Task Duration Spent Looking In Each Location

Figure 3-11 shows the Proportion of Task Duration Spent Looking at each Location Type by Task. This metric brings together number of glances and length of glances into a total time per location type and divides that by the task duration so that the values are expressed in terms of the proportions of each task-duration spent viewing each location type.

A caveat is appropriate before proceeding to the results. Proportions are most comparable when they relate to the same or similar durations. DWM task durations can differ considerably from task to task, especially among visual-manual tasks. Proportions mask duration differences even though those differences may be important. Proportions can be misleading when the task durations are substantially different. Therefore, caution is urged in the interpretation of this type of measure.

The Proportion of Task Duration Spent Looking at the Road Location metric, which refers to the test track in this chapter and is shown in blue on the graph, discriminated very well between tasks that are visual-manual and those that are primarily auditory-vocal. The mixed-mode tasks fell in between the visual-manual and auditory-vocal tasks. The Voice Dial resembled the visual-manual pattern more closely. Delta Flight Information resembled the auditory-vocal pattern more closely. The auditory-vocal tasks on this measure looked more like the Just Drive task, although even more time was spent looking at the road when performing them than when just driving (a proportion of 0.88 (88%) of a task's duration during auditory-vocal tasks versus ~0.83 (83%) of the Just Drive task). In contrast, during visual-manual tasks, time spent looking at the road dropped to between 0.34 (34%) and 0.61 (61%).

Proportions of Task Duration Spent Looking at Situation Awareness Locations (mirrors and speedometer, shown in pink) varied over a narrower range across all tasks, with the Just-Drive task showing a slightly higher proportion of time on mirrors than when drivers were engaging in an additional in-vehicle task. In Figure 3-12, data for just the mirrors are shown in a similar manner with glances at the speedometer removed for this analysis, using the same measure. However, the scale has been enlarged, so that the magnitude of the effects can be compared. Relative to the Just Drive task, visual-manual tasks led to a larger drop in mirror viewing, on average, than did auditory-vocal tasks. For Just Drive, mirrors were viewed for 14.3 percent of a task's duration. On average, for auditory-vocal tasks, it dropped slightly to 11 percent, and for visual-manual tasks, it dropped further to 8 percent. In the graph, these percentages are plotted as their corresponding proportions: 0.143, 0 .11, and 0.08.

Additional special analyses were also conducted using linear mixed-model analyses to examine whether breadth of scanning narrowed under higher-workload tasks regardless of type. The outcome of these analyses confirmed that for the test track data, there were significant differences (at $p<0.05$) in mirror scanning behavior between tasks classified as high and low workload for four of eight measures examined on glances to the mirror location. These measures were: number of glances to the mirrors, total glance time to the mirror location, percent (proportion) of time during task spent viewing the mirror location, and maximum glance duration to the mirror location.

Looking back at Figure 3-11, the *Proportion of Task Duration Spent Looking at Task-Related Areas* (shown in yellow) depended heavily on the nature of the task, and was primarily related to the visual-manual tasks, ranging from 0.29 to about 0.60. This measure for the visual-manual tasks provided an overall indication what percentage of the task period was spent looking at the task, perhaps an overall indicator of visual demand. If conceived of in this way, the ranks from high- to low-visual demand among the visual-manual tasks were: (1) Destination Entry, (2) Radio (Hard), (3) HVAC, (4) Radio (Easy), (5) Route Tracing, (6) Read (Hard), (7) Read (Easy), (8) CD / Track 7, (9) Map (Hard), (10) Map (Easy), (11) Coins, (12) Cassette, and then the mixed-mode tasks of (13) Voice Dial and (14) Delta Flight Information. Auditory-vocal tasks, though generating some task-related glances up to the headliner/visor and/or rearview mirror areas, approached proportions of 0.00. Particularly interesting was the fact that for some tasks, the proportion of time spent viewing the task exceeded that spent viewing the road—these tasks can be identified in the graph where the task-related yellow line is above the blue line for road glances. These tasks included HVAC, Radio (Easy), Radio (Hard), Destination Entry, Read (Easy), Read (Hard), Map (Hard), and Route Tracing. The miss rates for CHMSLs tended to be higher for visual-manual tasks for which time spent viewing the task was equal to or greater than time spent viewing the road (see Figure 3-13).

Overall, proportions of total glance time did not support straightforward results for visual-manual tasks. The same proportion was sometimes associated with CHMSL miss rates that differed by a factor of two. For example, Read (Hard), CD / Track 7, Map (Hard), Read (Easy), and Manual Dial - Home tasks all had proportions of approximately 0.45. Despite this, the CHMSL miss rates for the first two tasks were 16 percent and 19 percent, respectively, as opposed to miss rates of 30 percent to 33 percent for the latter three tasks. Conversely, different proportions were sometimes associated with nearly identical CHMSL miss rates. For instance, Map (Easy) (38% task-related total glance duration) was lower in proportion than Route Tracing and Radio (Easy) (about 50%). Despite this difference, these tasks had nearly identical CHMSL Miss Rates (33%).

There is some evidence for the notion that these measures based on proportion of time spent viewing specific areas (e.g., the task, and perhaps especially a ratio of the proportion viewing task to proportion viewing road), may provide an indication of visual demand. Other measures that

address number of glances or glance times rather than proportions may prove more useful. Such hypotheses may be interesting to explore in future research.

The measure of Proportion of Task Duration Spent Looking at the Road Location showed a high proportion of time spent looking at the road during auditory-vocal tasks. The data reported earlier in the chapter illustrated extended glance durations for the road location that were observed during auditory-vocal tasks. A question might arise as to what the state of attention was during these periods of extended looking. Was the driver attentive while looking at the road or was the driver in a period of inattentiveness, perhaps gazing steadily ahead but not really "seeing"? Were these prolonged glances indicators of a state of non-responsiveness to external stimuli? Extended roadway gazing might have been associated with being inattentive during auditory-vocal tasks. If so, then it was likely that drivers would have been non-responsive to events that were presented during tasks that were characterized by this type of eyeglance behavior. Figure 3-13 and Figure 3-14 were generated to examine this relationship.

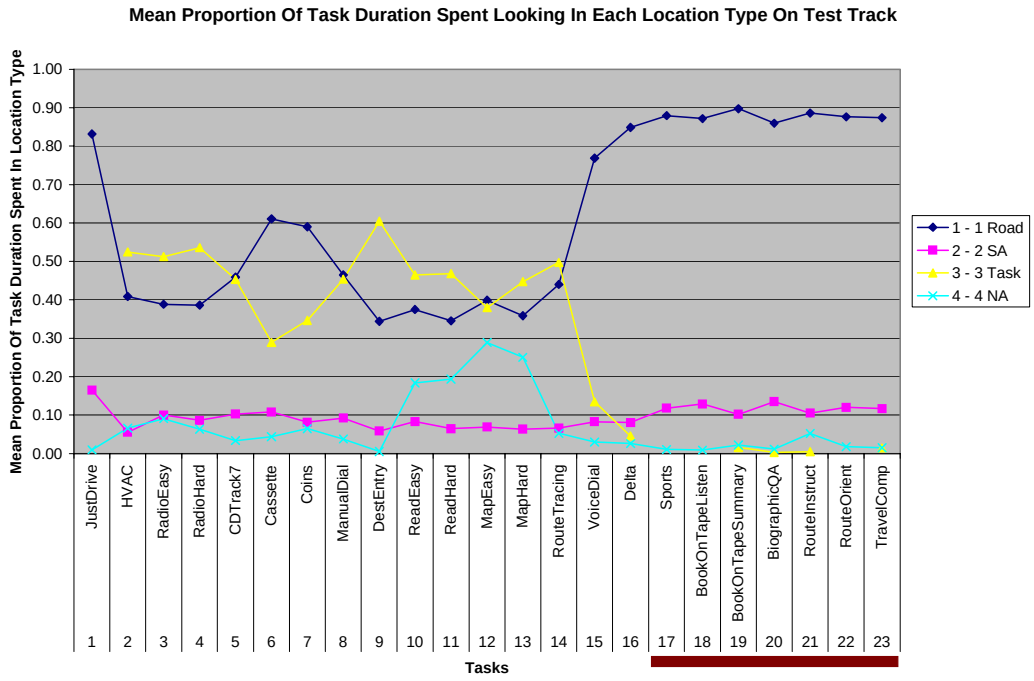


Figure 3-11. Track Mean Proportion of Task Duration Spent Looking in each Location Type by Task

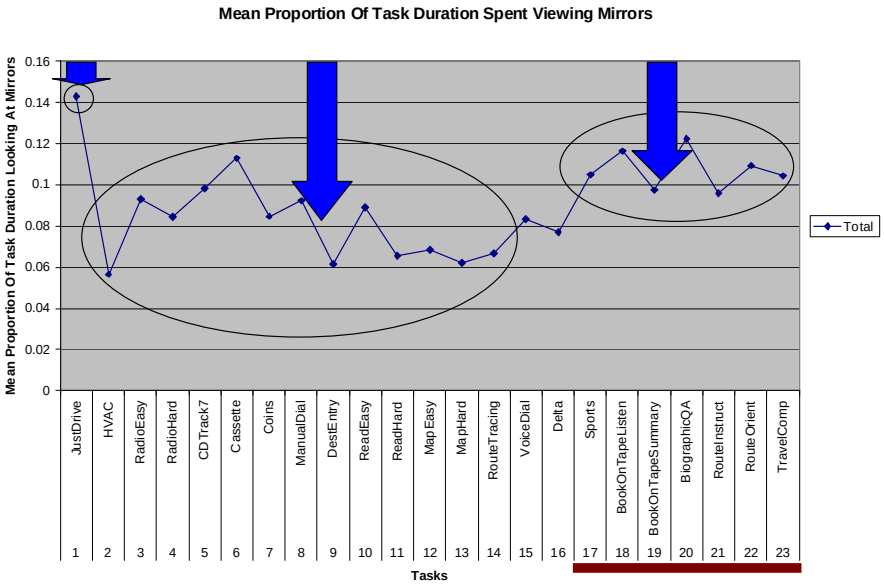


Figure 3-12. Track Mean Proportion of Task Duration Spent Viewing Mirrors (to compare with similar line for SA locations)

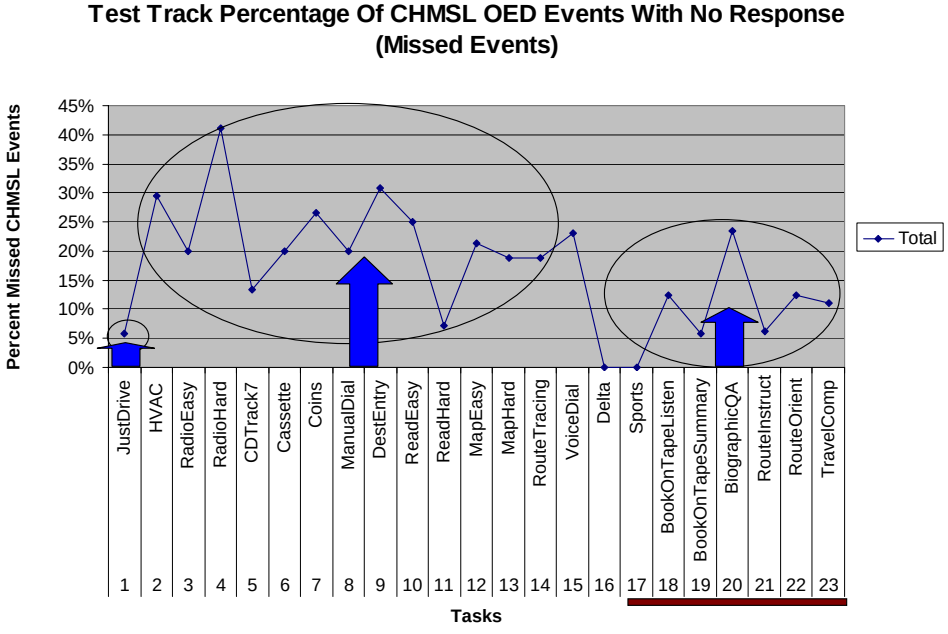


Figure 3-13. Percent CHMSLs Missed on the Track by Tasks (for comparison to glance patterns)

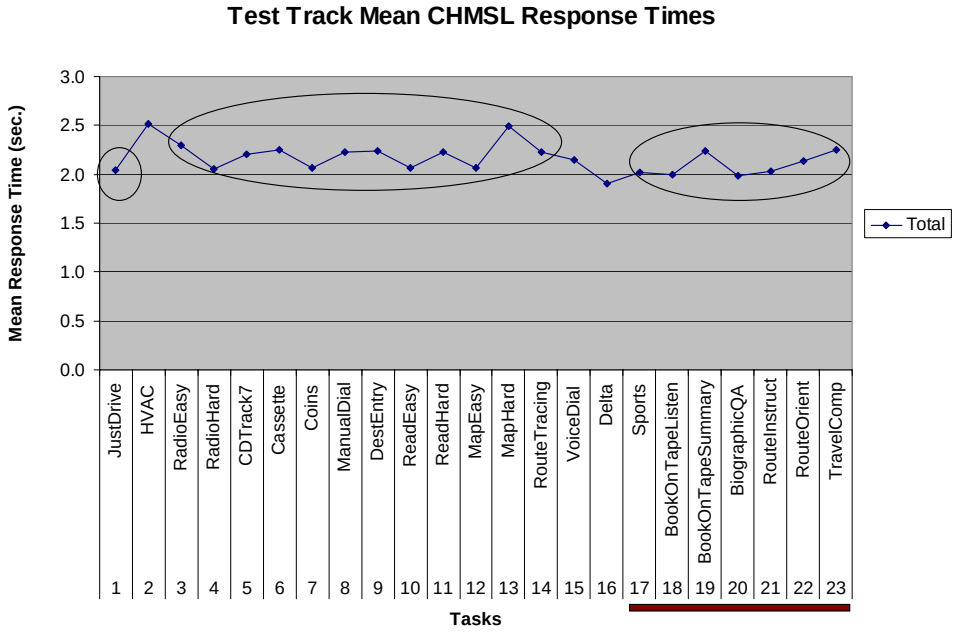


Figure 3-14. Response Times to CHMSLs on the Track by Tasks (for comparison to glance patterns)

If the long glance durations and concentration of gaze on the forward roadway were indications that drivers were inattentive to road during auditory-vocal tasks, then measures of attentiveness to event detection should indicate that higher percentages of events were missed during these auditory-vocal tasks. Specifically, Figure 3-13 should show an elevation in the percent missed CHMSL events on the Track. Figure 3-13 and others like it are based on data from the 18 research participants from whom eye data were reduced for the test track venue. These figures, therefore, represent only a sub-sample of the larger data set on event detection described earlier in this chapter.

The points in Figure 3-13 that were associated with auditory-vocal tasks showed a slight elevation in Percent Missed CHMSLS over Just Drive for some tasks and even lower miss rates than Just Drive for others. Averaging over auditory-vocal tasks, 10.4 percent of CHMSLs were missed, versus 6 percent for Just Drive, shown by the arrows in the figure. However, on average, there was an even greater elevation in percent missed CHMSLs for visual-manual tasks than for auditory-vocal tasks; 19.5 percent for visual-manual versus 10.4 percent for auditory-vocal, shown by the arrows in the figure. Similarly, if drivers were inattentive to the road during auditory-vocal tasks, Figure 3-14 should show clearly slower Response Times (RTs) associated with auditory-vocal tasks on the track. However, the pattern for Response Times to CHMSL events in Figure 3-14 showed very little difference between task types, on average. RTs to CHMSLs for Just Drive were 2.04 seconds on average, 2.05 seconds for visual-manual tasks, and 2.09 seconds for auditory-vocal tasks. Together, these results for Percent Missed CHMSLs and RTs to CHMSLs suggest that the concentration of gaze on the forward roadway observed in drivers performing auditory-vocal tasks on the test track was associated with only very subtle changes in attentiveness to CHMSL events, an increase in miss rate of between 0.5 percent and 9 percent and a decrease in miss rate for other tasks of 1.2 percent to 2.3 percent. It was much less pronounced than that produced by visual-manual tasks, affecting primarily miss rate.

The data on responsiveness to FVTS events is shown in Figure 3-15 and Figure 3-16 (both related to FVTS event detection. These data indicate that there was somewhat more interference from the auditory-vocal tasks on detection of these peripheral events, which appeared in the left outside mirror, but it was again less than that produced by the visual-manual tasks. Averaging across the auditory-vocal tasks shown in Figure 3-15, 45.43 percent of FVTS events were missed versus 22 percent for Just Drive. This compared with 63.54 percent missed FVTS for visual-manual tasks. These average miss rates are depicted by the arrows in the figure. While the level of inattentiveness was still considerably less for auditory-vocal than for visual-manual tasks, it was more distinct for FVTS events than for CHMSLs (occurring mostly for the three most difficult Auditory-Vocal tasks). It was also consistent with the findings in

Figure 3-12, indicating some reduced scanning of the mirrors during auditory-vocal tasks and even more reduced scanning of mirrors for visual-manual tasks. In Figure 3-16, response time data indicated that during Just Drive, participants responded to FVTS events within 2.55 seconds on average. For auditory-vocal tasks, the response times averaged 2.53 seconds. Response times for visual-manual tasks were more variable, averaging 2.09 seconds across the set, but with two subsets of visual manual tasks showing somewhat different patterns. The shortest task (HVAC) showed fast response times, though why is unclear. Similarly short RTs were associated with tasks that allowed drivers to hold materials in their hands and move it with their line-of-sight—Read (Easy), Read (Hard), Map (Easy), Map (Hard), and even Route Tracing. The other visual-manual tasks showed somewhat longer response times to detected FVTS events.

The results for responsiveness to LVD events were similar to those for CHMSLs, as shown in Figure 3-17 and Figure 3-18 (both related to LVD). For auditory-vocal tasks, which were associated with a concentration of gaze on the forward roadway, there was a slight elevation in Percent Missed LVDs when compared to Just Drive (16.6%, on average, versus 13% for Just Drive) but this slight elevation was again less than that seen for visual-manual tasks (33%) on average, though some of these had rates of missed LVDs for the methodological reason that they were too short for the event to even be detectable within the task's length. (See the Chapter 8, *Discussion and Recommended Toolkit*, for details on this point.) Arrows depict these average miss rates in the figure. Response times to detected LVD events (Figure 3-18), on average, were 5.25 seconds for Just Drive, 5.45 seconds for auditory-vocal tasks, and 5.90 seconds for visual-manual tasks, a pattern consistent with the miss rates.

Therefore, a hypothesis that very long glances to the forward roadway during the auditory-vocal tasks in this study may have indicated some level of inattentiveness, received mixed support from the data. Just Drive was associated with missed detections. Two auditory-vocal tasks had even lower missed detection rates than Just Drive. The remaining auditory-vocal tasks were associated with higher miss rates but a simple Biographical Q&A task had the highest miss rates. There is no clear pattern. The magnitude of the effects was much smaller than might have been expected, and attentiveness to events was higher during auditory-vocal tasks than during visual-manual tasks. Effects for peripheral FVTS events were consistent with the notion that scanning of the periphery was shed to some degree during auditory-vocal tasks, as the eyes concentrated on the forward roadway more during auditory-vocal tasks. However, for visual-manual tasks, it appeared that scanning of the periphery was shed to a much greater extent than for auditory-vocal tasks, so that drivers could look back and forth between in-vehicle activity and the forward roadway.

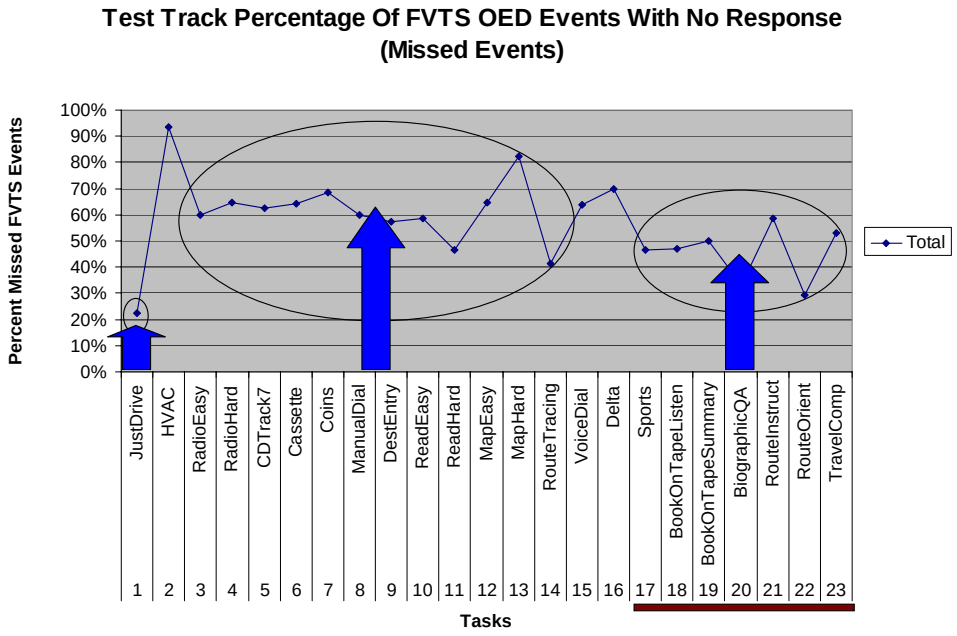


Figure 3-15. Percent FVTSs Missed on the Track by Tasks (for comparison to glance patterns)

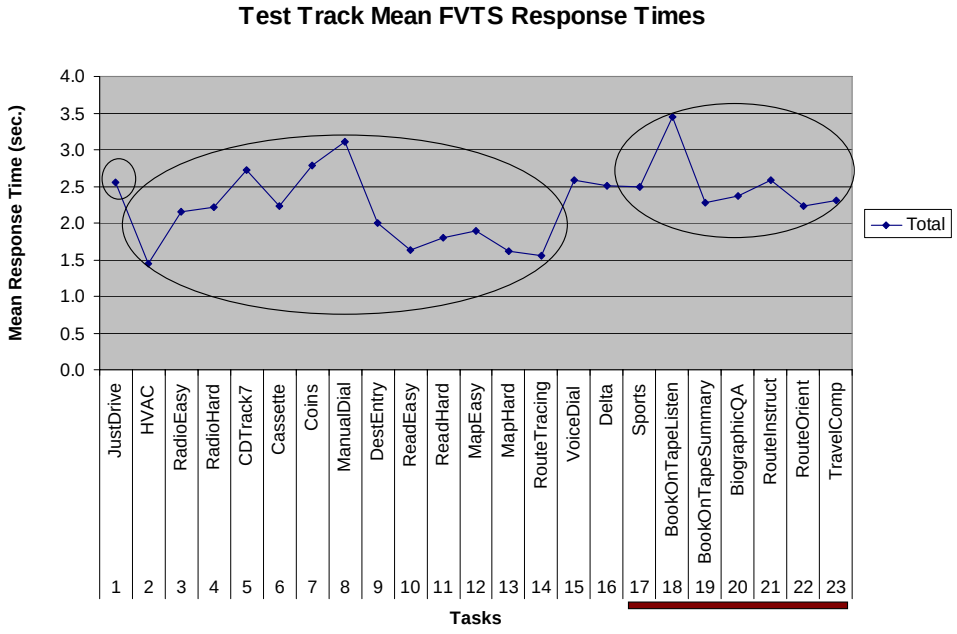


Figure 3-16. Response Times to FVTS Responded to on the Track by Tasks (for comparison to glance patterns)

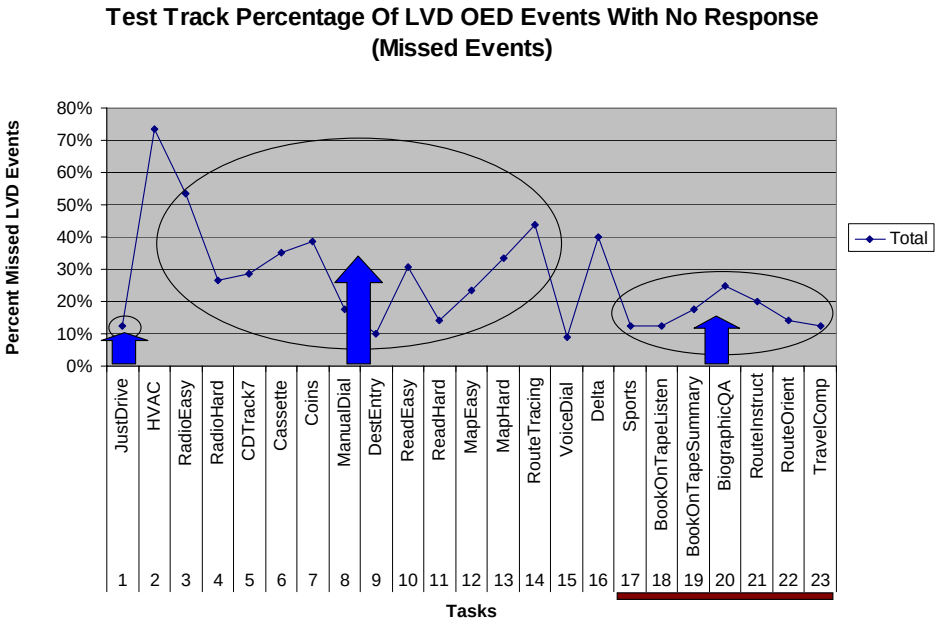


Figure 3-17. Percent LVDs Missed on the Track by Tasks (for comparison to glance patterns)

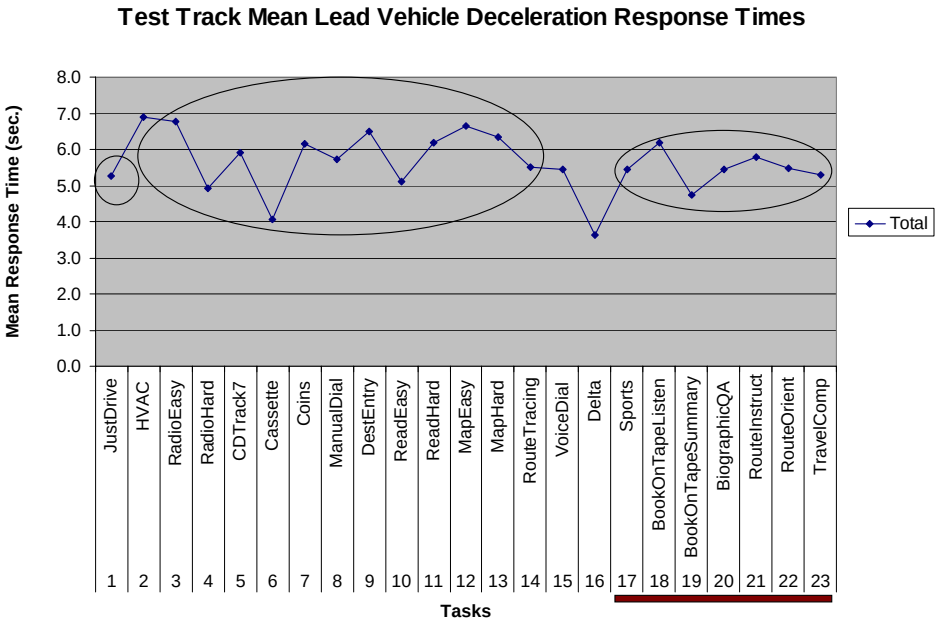


Figure 3-18. Response Times to LVDs Responded to on the Track by Tasks (for comparison to glance patterns)

In summary, there were multiple effects of in-vehicle tasks on eyeglance behavior. Eyeglance metrics showed distinct patterns for different types of tasks—Just Drive versus concurrently performing an auditory-vocal task or concurrently performing a visual-manual task. The Just Drive task was distinguished by patterns in which drivers looked at the road about 83 percent of the time and scanned their mirrors about 14.3 percent of the time. Glances on the road were about 8 seconds long, on average.

Auditory-vocal tasks showed a somewhat similar pattern, although drivers gazed at the forward roadway somewhat more (88%), using longer gazes (9 to 16 seconds, on average), and scanned their mirrors somewhat less (11%). The miss rate for event detection was slightly elevated over just driving for auditory-vocal tasks for CHMSL and LVD events (showing an increase of ~4% for CHMSL and LVD events). However, two auditory-vocal tasks were associated with miss rates below Just Drive alone. The miss rate for event detection was somewhat more for peripheral FVTS events (showing an increase of ~23%), although event detection was less affected by auditory-vocal tasks than by visual-manual tasks.

Visual-manual tasks showed a different pattern in which drivers looked at the forward roadway much less, viewing the road only 34 to 61 percent of the time during a task and using glance durations on the road that were less than 2 seconds long, on average. This reduction in glances to the road was made largely in order to view task-related areas required for performing the in-vehicle activity (viewing the task 29 to 60 percent of the time during its length). For visual-manual tasks, glances tended to cycle frequently back-and-forth between the task and the roadway locations, and glance-rate measures proved to carry interesting information reflect this. Visual-manual tasks led to a more pronounced reduction in mirror-scanning (to 7%) and were associated with higher rates of missed events, although this was sometimes due to a methodological constraint for LVDs. Increases in miss rates over Just Drive were approximately 14 percent for CHMSLs, 20 percent for LVDs, and 42 percent for FVTS events, on average.

3.4.2 Event Detection and Glance Patterns Relationships

Early exploratory analyses of glance duration measures revealed that differences in glance patterns were present when events (such as CHMSLs and FVTSs) had been responded to by drivers. In particular, these initial analyses indicated that glance durations tended to shorten on trials in which events had been detected and glance frequencies to certain locations tended to increase. This suggested that a change in glance patterns had occurred; one in which more eye movements were occurring with shorter gazes in between. It provided a clue that the occurrence of an event may have induced more scanning of the visual field than had been occurring previously.

Early analyses of the data included an examination of selected time series plots of the data. An example of such a plot is shown in Figure 3-19. An enlargement of the area of interest is shown in Figure 3-20. It illustrates an increase in glances to the mirror following a driver's response to an event in the trial. It was hypothesized that an event of this type may serve as an attentional interrupt, which serves to attract additional scanning to increase situational awareness surrounding a possible threat or risk.

Specifically, Figure 3-19 shows the data for one participant who detected and responded to a CHMSL event. The participant's response to the CHMSL is depicted by the line showing the point at which "Driver Switch" was depressed (shown in navy blue). The glance location type depicted by the black line is that of mirrors. The frequency of looks to the mirrors increased immediately following the response by the driver to the CHMSL event.

Figure 3-20 shows the same data for the same participant, but with enlargement. The participant's response to the CHMSL is depicted in this plot by the line showing Driver Switch being depressed (red line in this plot). As noted, the frequency of looks to the mirrors increased immediately following this response by the driver to the CHMSL event.

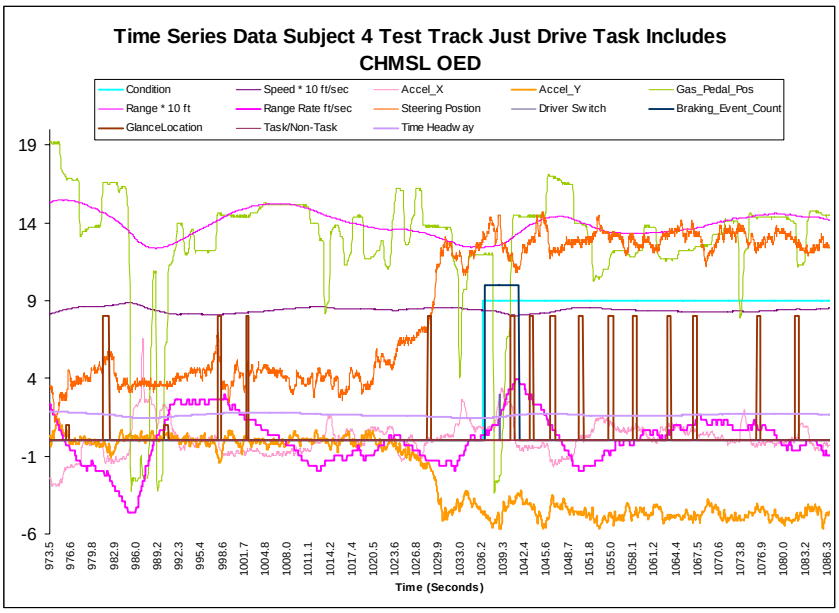


Figure 3-19. Time Series Plot Depicting Change in Glance Patterns Following Occurrence and Detection of Visual Event During Trial

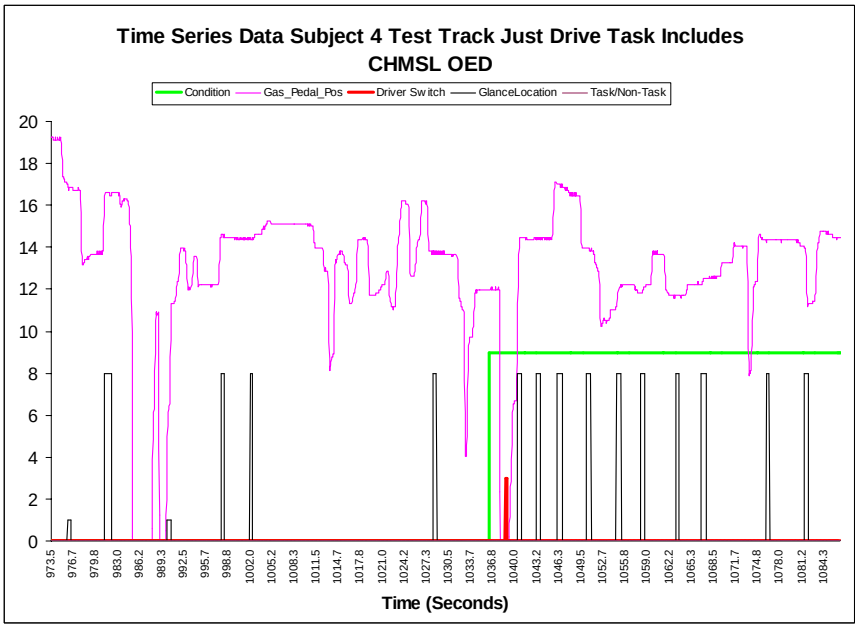


Figure 3-20. Time Series Plot Depicting Change in Glance Patterns Following Occurrence and Response to Visual Event During Trial (with simplification and enlargement of the area of interest)

Linear Mixed-Model Analyses were conducted to provide formal statistical tests of the hypothesis that events, like CHMSL illumination or LVDs, may function as attentional interrupts that serve to attract additional scanning which would increase SA surrounding a possible threat or risk. These Linear Mixed-Model Analyses were separately conducted on each type of event detection (i.e., CHMSL, FVTS, and LVD). Emerging from these Linear Mixed-Model Analyses were significant main effects of Task, Location, and Detect Event (or Event Response), which were qualified by interactions of Location by Detection and, in some instances, Task by Detection. See Table 3-4. Though the Linear Mixed-Models table uses the factor name Detect, or Detection, in this report it will be referred to as Event Response, since it refers to whether the driver did or did not respond to the event that was presented during driving.

In addition, it is important to understand that in these analyses it was not possible to separate the glances occurring within a task in terms of whether they occurred before or after an event. To do so would have necessitated recoding the database, which was not feasible within the timeframe remaining for the project. Thus, all the glances made within a task had to be treated in the same way, and all were included in the analysis. The consequence of this is that in order for the analysis to detect an effect of events on glance metrics, the effect would have to be sizeable—since its effects would only be exerted on glances occurring after the event occurred, and these glances will be averaged in with all other glances occurring during the task, thus diluting the effect for trials on which a detection occurred.

Table 3-4. Linear Mixed-Models Effects for Analyses of CHMSL, FVTS, and LVD Detection Responses and Their Effects on Eyeglance Behavior

Test Track	Effect	# Glances	Total Dur	Max Dur	Min Dur	Mean Dur	Medn Dur	St Dev Dur	Glance Rt
CHMSL	Task	*	*	*		*	*	*	*
	Location	*	*	*				*	*
	Detect CHMSL	*							
	Task * Detect			*				*	
	Locat* Detect	*	*	*		*	*	*	*
Test Track									
FVTS		# Glances	Total Dur	Max Dur	Min Dur	Mean Dur	Medn Dur	St Dev Dur	Glance Rt
	Task	*	*	*		*	*	*	*
	Location	*	*	*	*	*	*	*	*
	Detect FVTS			*	*	*	*	*	*
	Task * Detect	*		*		*	*	*	
	Locat* Detect	*	*						*
Test Track									
LVD		# Glances	Total Dur	Max Dur	Min Dur	Mean Dur	Medn Dur	St Dev Dur	Glance Rt
	Task	*	*	*		*	*	*	*
	Location	*	*	*	*	*	*	*	*
	Detect LVD					*			
	Task * Detect					*			*
	Locat* Detect	*	*	*			*	*	*

The Linear Mixed-Model analyses confirmed that although there were significant main effects of Task, Location (road, SA, task, and NA), and Event Response (yes/no, labeled “Detect” in the Linear Mixed-Models table), there were also statistically significant interactions involving these variables. Of particular interest were the statistically significant interaction effects of Location by Detect (Event Response) on multiple metrics, as indicated by the asterisks in the table above (asterisks designate effects significant at the $p \leq 0.05$ level). Green highlighting identifies effects that have been explored graphically as well as statistically. Graphs for the key highlighted interactions are included in what follows to illustrate the key effects.

In brief, these graphs suggest that when an event occurs and is responded to, eyeglance behavior changes such that:

- For CHMSL events:
  - Durations of glances decreased slightly for all locations exception SA.
  - Rate of glancing increased slightly to road and situation awareness areas (mirrors).
- For LVD events:
  - Durations of glances to the road lengthened.
  - Rate of glancing decreased to task-related and situation awareness areas.
- For FVTS events:
  - Durations of glances decreased.
  - Rate of glancing to road and situation awareness areas increased.

Changes to glance durations interacted with Task Type and were more pronounced for Just Drive and auditory-vocal tasks than for visual-manual tasks (which usually showed a different pattern).

3.4.2.1 Number of Glances (as affected by Location by Event Response)

Figure 3-21, Figure 3-22, and Figure 3-23 show the significant Location by Event Response interaction on the Number of Glances metric. Each plotted point is the average of all 23 tasks. As can be seen, there was a large increase in the number of glances to the road and SA location types when any of the events had been detected and responded to (but little or no increase in glances to task-related areas). For the LVD event (Figure 3-23), glances to the task also increased somewhat following detection and response to an LVD event. One possibility for this is that LVD events may, to a greater extent than the other events, modify or interrupt task performance when they are detected, insofar as the magnitude of the driver's manual response is concerned. The driver must respond by removing the foot from the accelerator pedal and then tap the brake. This may interrupt or suspend manual activity on the in-vehicle task for a moment, and then that task activity may resume. This may be reflected in an increase in glances to the in-vehicle task following the detection of the LVD.

In Figure 3-21 and others like it, each plotted point was obtained by averaging across all glances to a location, such as to the road, SA, or task) and across all tasks. Visual-manual tasks were shorter and were associated with more missed events or non-responses, which are plotted in blue on these graphs. Visual-manual tasks thus contributed more data to the "blue" or "Miss" (no detect) points than did the auditory-vocal tasks. Auditory-vocal tasks contributed more data to the "pink" or "Detect" points than did the visual-manual tasks. Thus, there is the possibility that type of task also interacts with Location and Event Response. This relationship will be graphically depicted in subsequent figures starting with Figure 3-30.

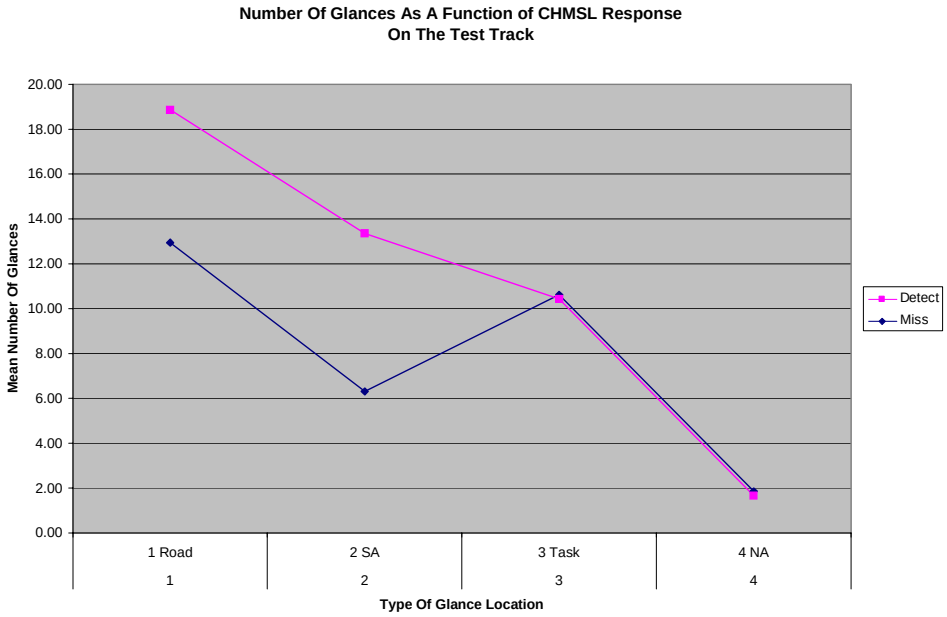


Figure 3-21. Effect of Response to CHMSLs on Number of Glances by Glance Location

Note: Shows a large increase in the number of glances to the road and SA location types when a CHMSL has been responded to (but not to task-related areas)

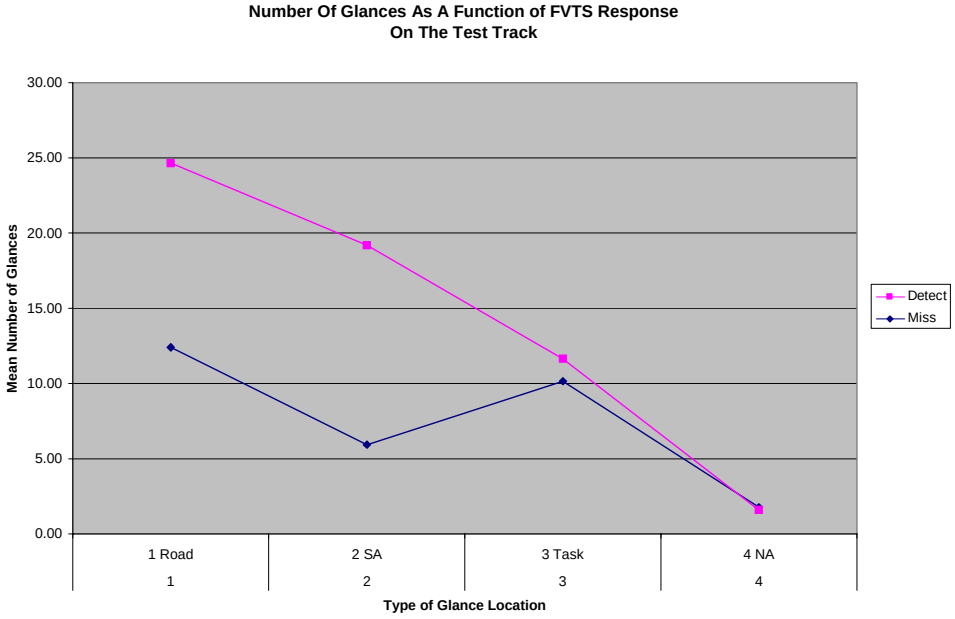


Figure 3-22. Effect of Response to FVTS Events on Number of Glances by Glance Location

Note: Shows a large increase in the number of glances to the road and SA location types when a FVTS has been responded to, but only a very small, almost negligible effect on number of glances to the task.

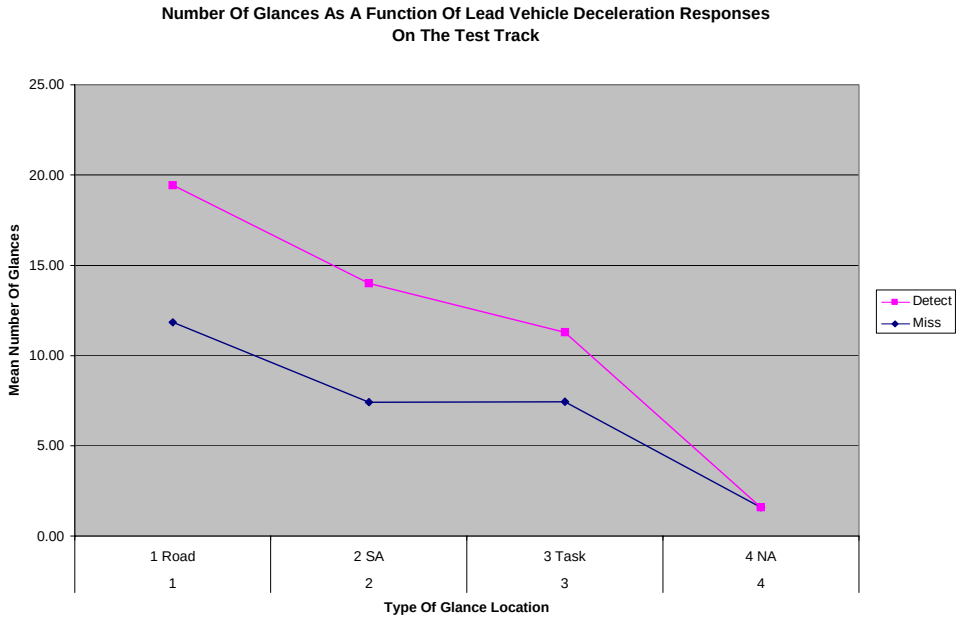


Figure 3-23. Effect of Response to LVD Events on Number of Glances by Glance Location

Note: Shows a large increase in the number of glances to the road and SA location types, and also a smaller increase in glances to task-related areas

3.4.2.2 Glance Duration (as Affected by Location by Event Response)

Figure 3-24, Figure 3-25, and Figure 3-26 show the interaction of Location by Event Response, but for metrics related to Glance Duration. The interaction indicates that for CHMSLs there is a small but reliable decrease in the duration of glances to all locations (road, task, and NA) except for those related to situation awareness (mirror and speedometer checks, which are already very short on average). This is shown in Figure 3-24, which depicts Mean Glance Duration, a metric on which this interaction was significant (as well as on Median Glance Duration). For comparison, Figure 3-25 shows a similar pattern for FVTS events, though it was not significant in the Linear Mixed-Model Analyses. Figure 3-26, shows the interaction for LVD events and it is quite different from the pattern for CHMSL and FVTS events. In the case of LVDs, glances to the road area increase in length (rather than decrease). Such a response would allow drivers to acquire more information about a decelerating vehicle over time, and so would appear to represent an appropriate adaptation to the detected LVD. Conversely, longer glances to the road may have resulted in more LVD detections. Some evidence for this comes from task differences. Auditory-vocal and Just Drive tasks were generally much longer than visual-manual tasks. Auditory-vocal tasks contributed to a larger proportion of the trials averaged together for the "Detect" data point. On the other hand, visual-manual tasks were shorter and had more "Miss" (missed detections) for perceptual reasons discussed later in this report. This alone could account for the differences in single-glance durations to the road.

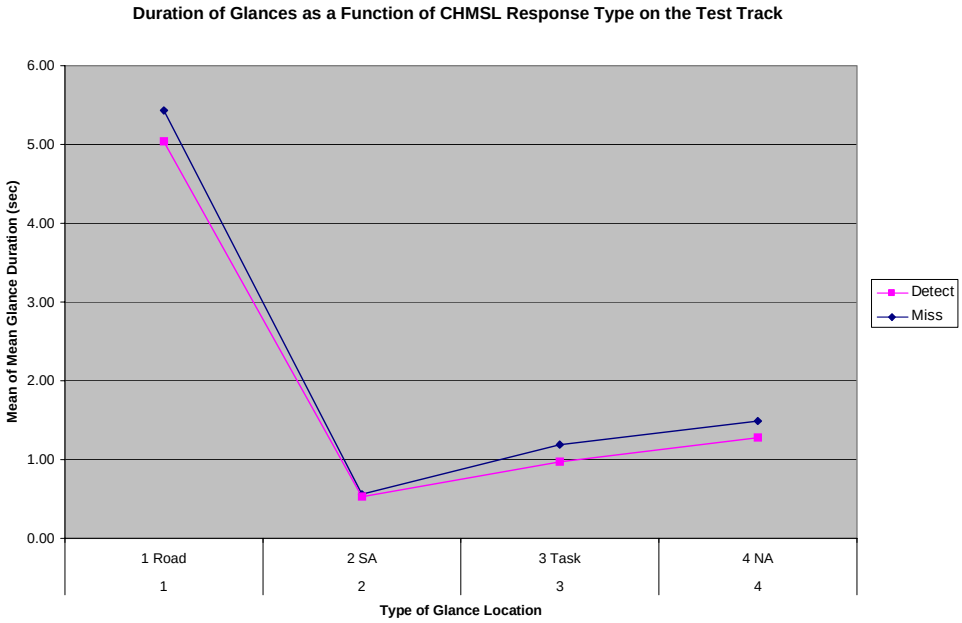


Figure 3-24. Duration of Glances (Based on Mean Glance Duration) as a Function of CHMSL Response Type (Detect/Miss) and Type of Glance Location for the Significant Location by Event Response Interaction

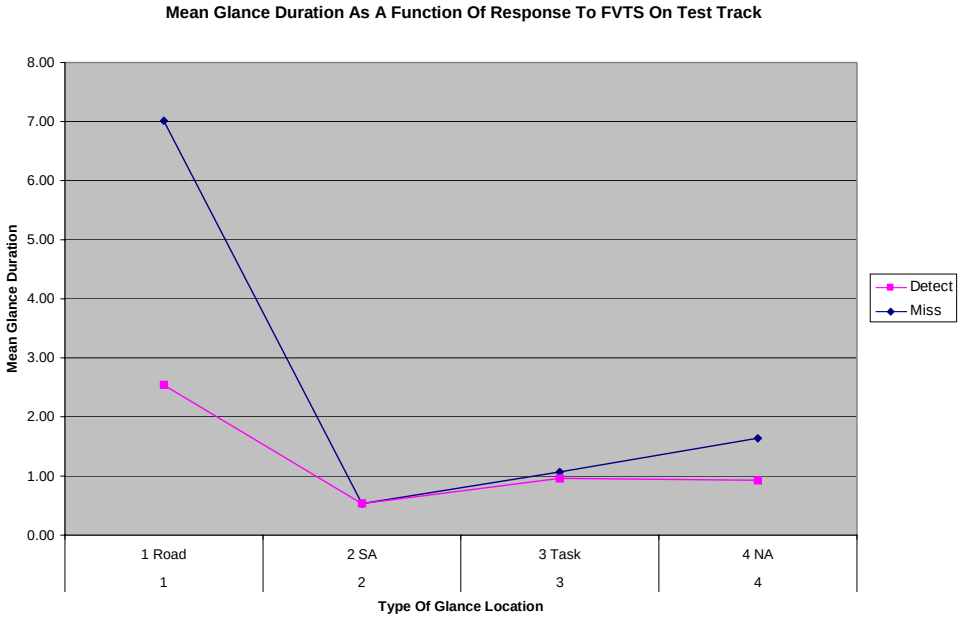


Figure 3-25. Duration of Glances (Based on Mean Glance Duration) as a Function of FVTS Response Type (Detect/Miss) and Type of Glance Location

Note: Location by Event Response interaction was non-significant for FVTS events, though the pattern was consistent with CHMSL

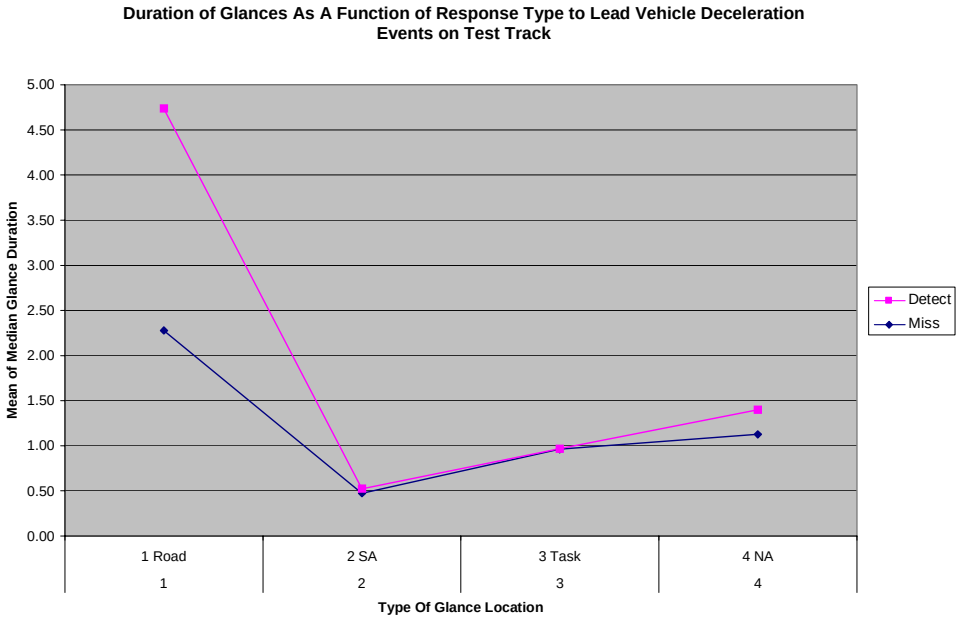


Figure 3-26. Duration of Glances (Based on Mean Glance Duration) as a Function of LVD Event Response Type (Detect/Miss) and Type of Glance Location for the Significant Location by Event Response Interaction

3.4.2.3 Glance Rate (as affected by Location by Event Response)

Figure 3-27 through Figure 3-29 show the Location by Event Response Interaction for Glance Rate. Changes in number of glances and durations of glances following event detection and responses, as well as total task duration, translate to changes in rates of glances per second. Figure 3-27 depicts the interaction for CHMSL events and Figure 3-28 depicts it for FVTS events. They both indicate that larger increases in glance rate occur for the situation awareness locations than for task-related or road locations (though there are small increases in glance rates to these areas as well). The increase in glance rate to the road is larger for the FVTS than for CHMSL. Figure 3-29 depicts the interaction for LVD events. Again, the pattern is quite different. There is virtually no change in the glance rate to the situation awareness category, but there is a decrease in glance rate to the road and task-related areas, which corresponds with increasing durations of glances to the road seen in Figure 3-26.

Together these patterns may suggest that driver glance patterns are modified in response to which events are detected and responded to, and that they are modified in a way that is specific to the event. These results suggest that drivers adapt their visual scanning in a way that is perhaps tailored for updating their awareness of current traffic and road conditions relative to the specific event they are responding to, and the types of risks it may represent to them. For example, when a CHMSL illuminates in front of them, drivers may habitually check mirrors to determine whether a lane change may be possible should the vehicle in front suddenly stop. When a follow vehicle signals a turn, looks to the mirror may increase to ascertain whether an overtaking maneuver will be initiated. And even though these conditions were not really relevant under the experimental conditions of the platoon methodology used in this experiment, scan patterns learned over years of driving may nonetheless be triggered by the stimulus events used in the study.

On the other hand, it may be that the effects observed here were in some way unique to the event detection methodology employed in the experiment. For example, while a driver may habitually check mirrors in case a sudden stop by a lead vehicle requires an evasive lane change maneuver, the likelihood of this was reduced on the test track by (1) extremely light traffic and (2) no sudden hard braking for task after task, which might have been expected to cause drivers to learn that hard braking was unlikely, purportedly a driver-expectation that is a common contributor to rear-end crashes. If drivers in the study in fact did not expect hard braking, and were not changing their scanning patterns due to learned responses that are adaptive for driving, then perhaps they changed their scanning merely to detect events that they expected in the experimental paradigm. However, this explanation cannot account for a change in glance patterns after the detection of an event, since only one event per task was presented for detection on the test track. Thus, there would have been no point to changes in glance durations or increased scanning of road and mirror locations following detection of an event for experimental purposes, since it would not have improved event detection performance during the task.

Therefore, the first explanation offered seems the more plausible—that the stimuli used in this experiment triggered scan patterns learned over years of driving. However, confirmation of these findings through future experimental work would be desirable.

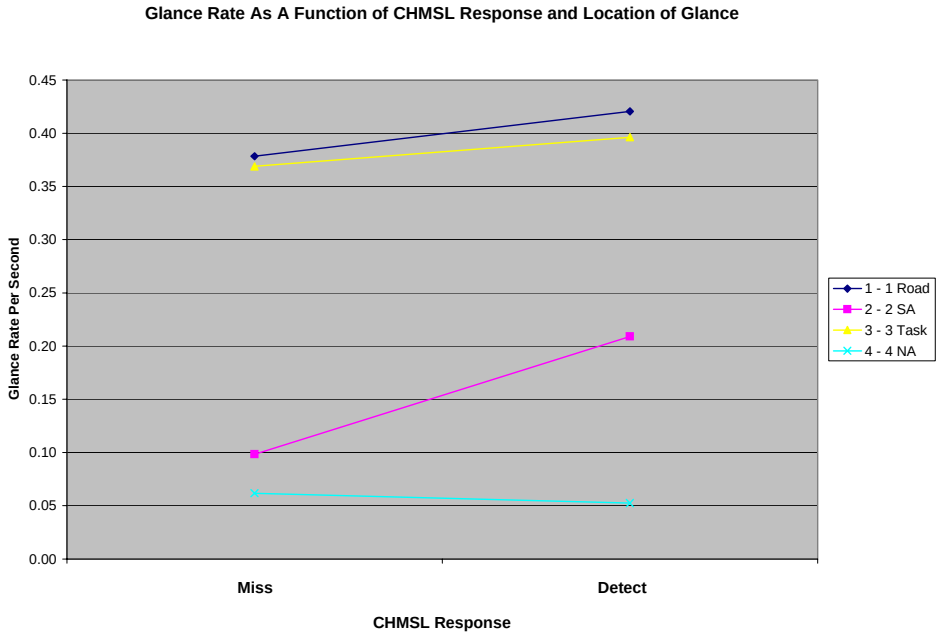


Figure 3-27. Interaction of Location by Event Response for CHMSL Events on the Glance Rate Metric

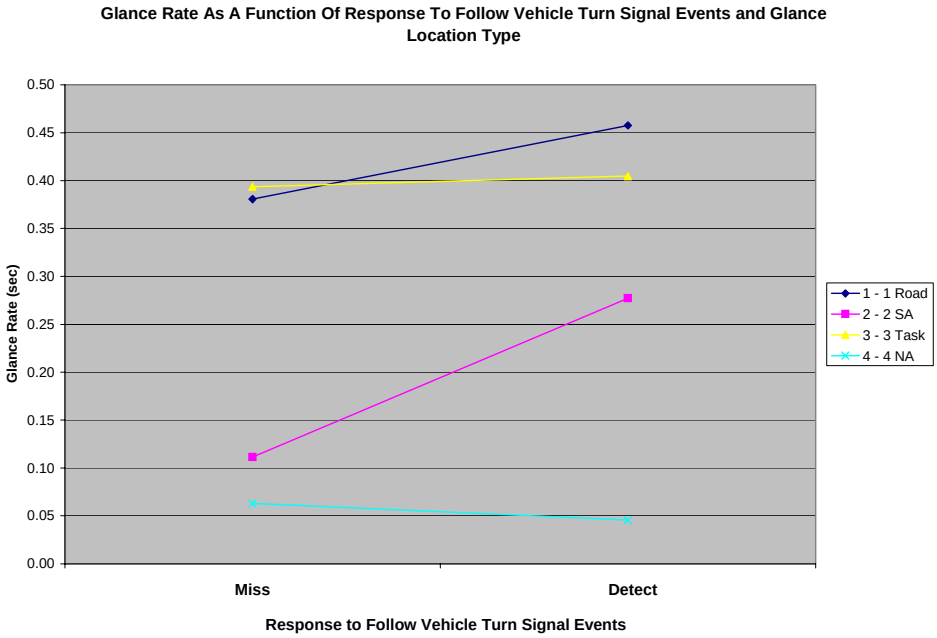


Figure 3-28. Interaction of Location by Event Response for FVTS Events on the Glance Rate Metric

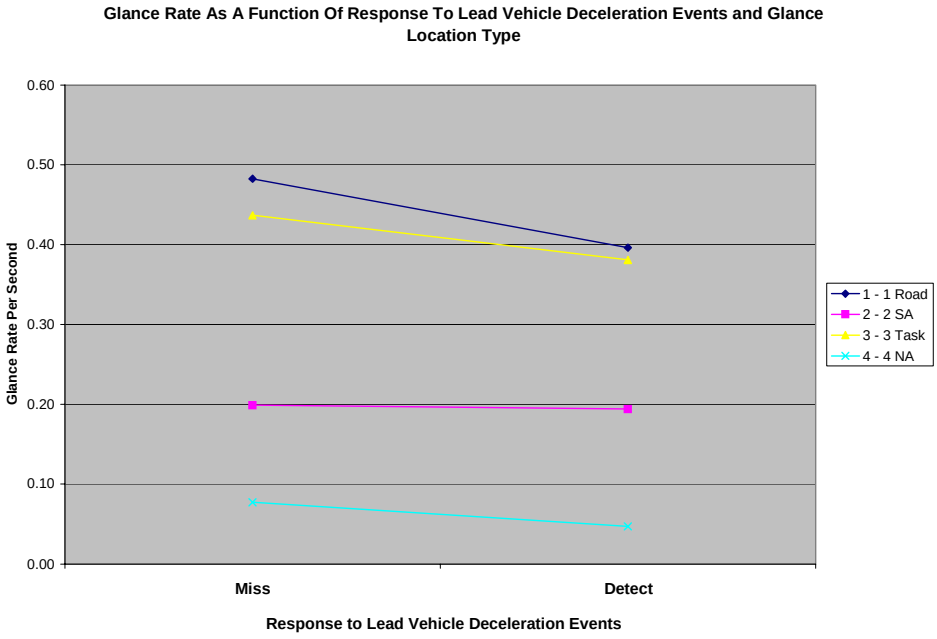


Figure 3-29. Interaction of Location by Event Response for LVD Events on the Glance Rate Metric

3.4.2.4 Glance Duration to All Locations Combined (in the Interaction of Task by Event Response)

Figure 3-30, Figure 3-31, Figure 3-32, and Figure 3-33, show the interaction of Task by Event Response on the metric of Glance Duration (averaged across all location types). In these figures, for each task each blue or pink point was obtained by averaging across all glances to all locations during that task. Furthermore, the set of data is based upon only the 18 participants from whom eye data were reduced from the test track venue. When these data are decomposed by task and then in terms of whether an event was detected or not, the data are sparse in some cells. For example, there are some cells in which no missed event-detections occurred. In those instances, a point will be missing from the plots.

The interaction for CHMSLs is illustrated in Figure 3-30 and Figure 3-31. In this interaction, it can be seen that the decrease in glance durations following detection and response to the CHMSL events is confined to a small set of tasks, primarily Just Drive, Read (Hard), Book-on-Tape Summarize, Biographical Q&A, Route Instructions, Route Orientation, and Travel Computations. The majority of these tasks are auditory-vocal tasks typically characterized by long glances at the road during task performance. These glances shorten following response to a CHMSL event and thus, there would be more of them. However, the interaction for CHMSLs is also due to the fact that for other tasks, most of the visual-manual tasks and two auditory-vocal tasks, the pattern is different. For the visual-manual tasks, except Read (Hard), there is virtually no change in Mean or Maximum Glance Duration as a function of having detected and responded to the CHMSL event. For the two auditory-vocal tasks, Sports Broadcast and Book-on-Tape Listen, there was an increase in Maximum Glance Duration across all location types. However, even though glance durations shorten for most of the auditory-vocal tasks following detection of a CHMSL event, they remain longer than for visual-manual tasks by a factor of three or more in most cases. This was likely due to the fact that most of the glances for auditory-vocal tasks were to the road location and longer versus split between the task and road and hence shorter.

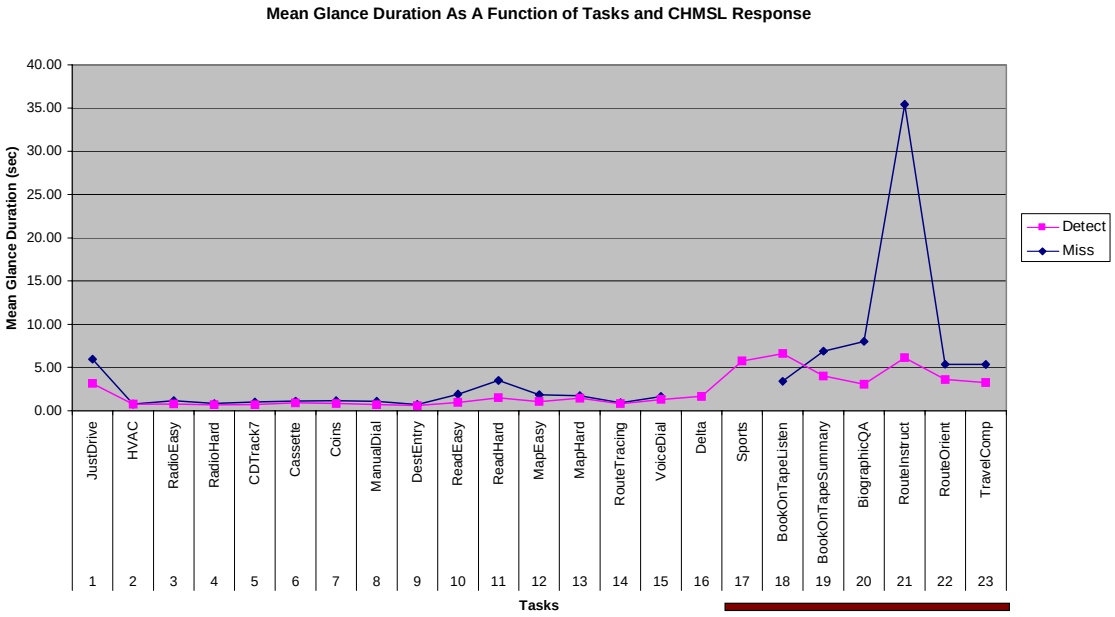


Figure 3-30. Non-significant Task by Event Response Interaction for CHMSL Events on the Metric of Mean of Mean Glance Durations

Note: Shown for comparison with other patterns.

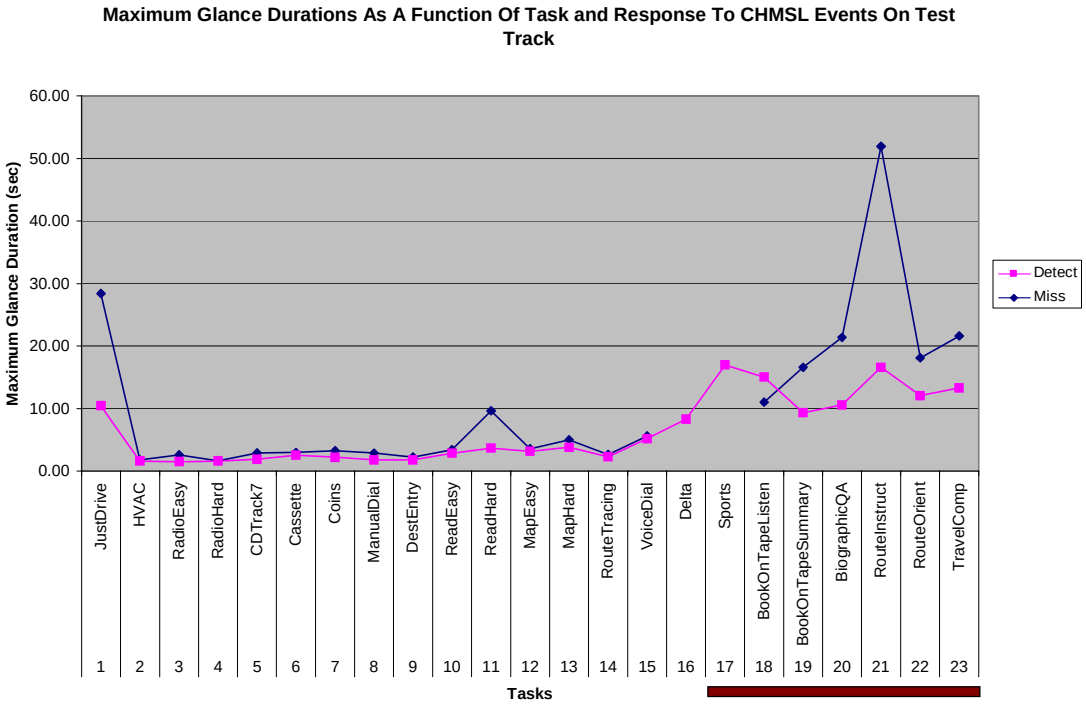


Figure 3-31. Significant Task by Event Response Interaction for CHMSL Events on the Metric of Mean of Maximum Glance Durations

Figure 3-32 shows the interaction for the FVTS events and the decrease in Mean Glance Duration across all location types as a function of "Detect" versus "Miss" (no detect) trials. Significant differences responsible for the results in Table 3-4, are confined to a small subset of tasks (Just Drive and auditory-vocal tasks). Visual-manual tasks show little or no change in mean glance duration as a function of FVTS event detection and response. Following the detection of an FVTS event, the mean glance duration, averaged across all locations, for Just Drive and auditory-vocal tasks more closely resembles that for visual-manual tasks, around two seconds or less. However, auditory-vocal tasks are as higher as or higher than any visual-manual task. This figure suggests that the effect on Glance Durations is not due just to task length, since Book-on-Tape Summarize was a short auditory-vocal task, only about 20 seconds versus approximately 2 minutes for the others, and still demonstrated the drop in glance durations for trials on which an event was detected.

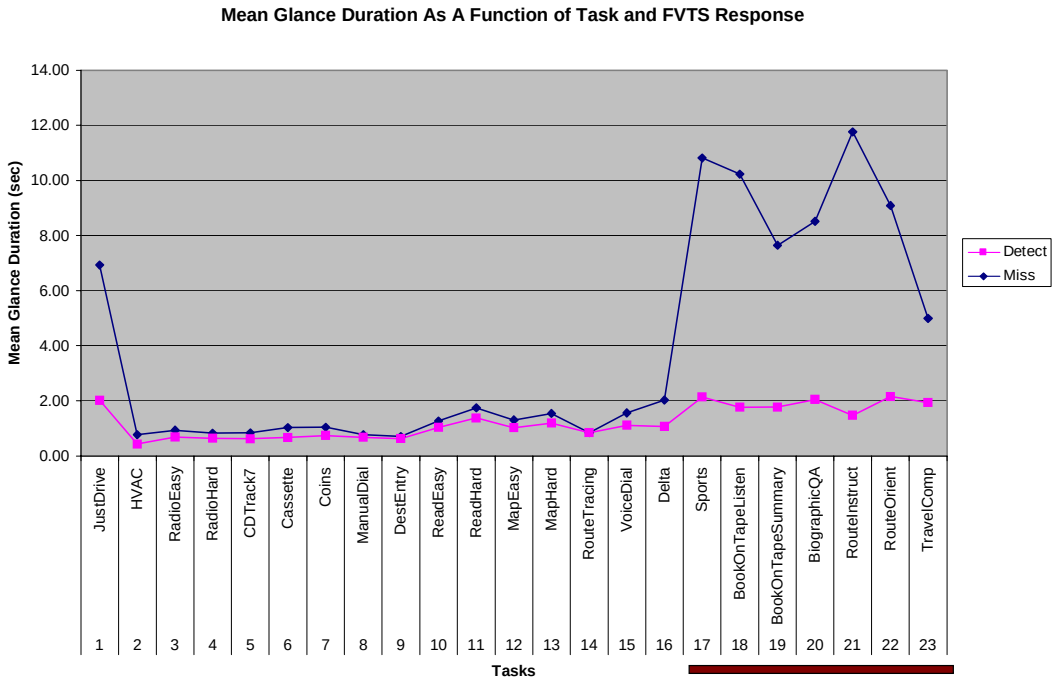


Figure 3-32. Significant Task by Event Response Interaction for FVTS Events on the Metric of Mean of Mean Glance Durations

Figure 3-33 depicts the significant interaction of Task by Event Response for LVD responses on the metric of Mean Glance Duration. It also indicates that the changes to glance duration primarily occurred on a subset of tasks, which were auditory-vocal in nature, along with Just Drive. Consistent with prior results, the pattern for LVD showed that glance durations lengthened—and this lengthening occurred on the auditory-vocal tasks of Sports Broadcast, Book-on-Tape Listen, Book-on-Tape Summarize, Biographical Q&A, and Route Instructions and for the task of Just Drive. It is likely that this effect can be attributed to longer glances to the road, based on prior results. Glance durations decreased for Route Orientation, Travel Computations, and Read (Hard). Visual-manual tasks, besides Read (Hard), showed little change in glance durations.

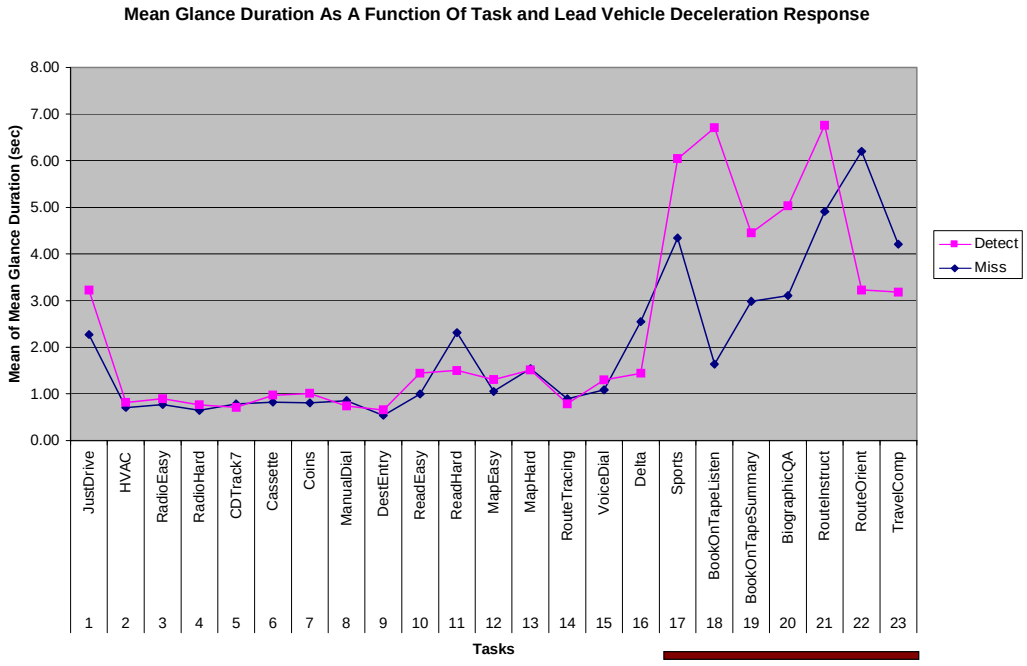


Figure 3-33. Significant Task by Event Response Interaction for LVD Events on the Metric of Mean of Mean Glance Durations

3.4.2.5 Glance Rate to All Locations Combined (in the Interaction of Task by Event Response)

Figure 3-34, Figure 3-35, and Figure 3-36 show the significant interaction of Task by Event Response for the Glance Rate metric for each of three event types. The interaction for CHMSL events is shown in Figure 3-34. Glance rates, averaged across all location types, increased following detection of a CHMSL for nearly all tasks, with a few exceptions—HVAC, Map (Hard), Route Tracing, and Book-on-Tape Listen. The interaction for FVTS events is shown in Figure 3-35. Here, glance rates also increased following detection of the event; this time for all tasks, though the increase for Route Tracing was negligible and the increases for visual-manual tasks tended to be smaller than for auditory-vocal tasks. For visual-manual tasks, an increased glance rate might be hypothesized to indicate continued attention to the task, with increased rates of scanning between roadway, task, and mirrors. For auditory-vocal tasks, an increased glance rate might be hypothesized to indicate something different—a shift of attention away from the task and to the situation. An increased glance rate resulted from less steady gazing at the roadway, and more scanning of road and mirrors, which is what leads to the hypothesis that attention may have shifted way from the auditory-vocal task and to the situation.

The interaction for LVD events is shown in Figure 3-36. Results showed that Glance Rate decreased for the "Detect" responses relative to the "Miss" responses for visual-manual tasks (plus Voice Dial). Glance rate increased for the "Detect" responses relative to the "Miss" responses for auditory-vocal tasks, Just Drive, and the mixed-mode task of Delta Flight Information. For visual-manual tasks, this pattern would be consistent with some type of reduced scanning between task and roadway locations in order to attend to the LVD event. The LVD event unfolded relatively slowly and may have required longer road glances to appreciate the change or rate-of-change in separation or lead vehicle visual angle. This same reduction was not

seen for CHMSL and FVTS events. For auditory-vocal tasks, though, it is the opposite pattern that would indicate a shift of attention to event monitoring. Namely, an increase in glance rate would indicate that a shift from steady gazing to active scanning of the forward roadway and mirrors during auditory-vocal tasks.

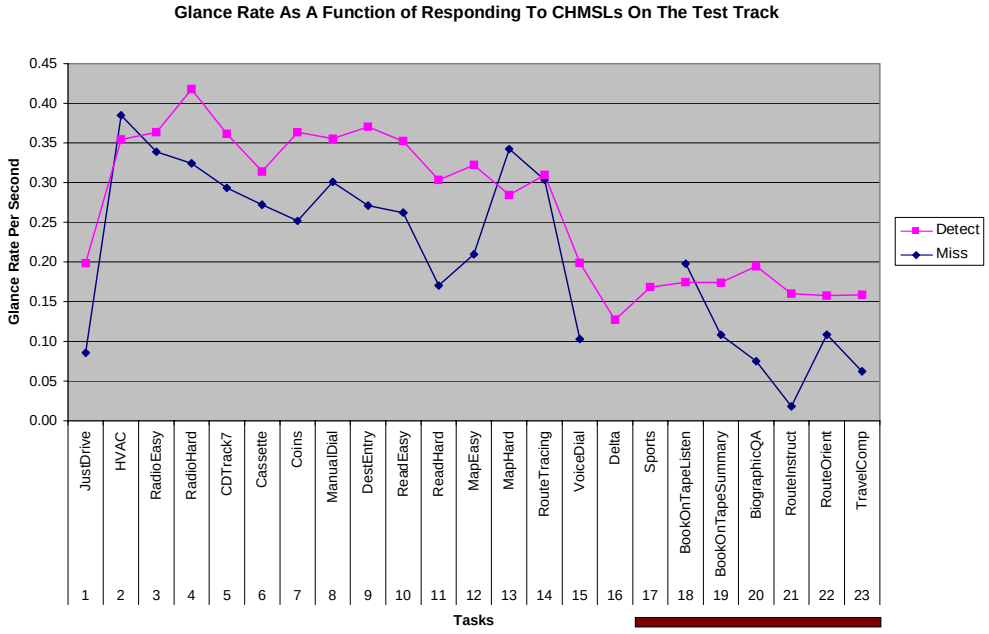


Figure 3-34. Task by Event Response Interaction for CHMSLs on the Glance Rate Metric

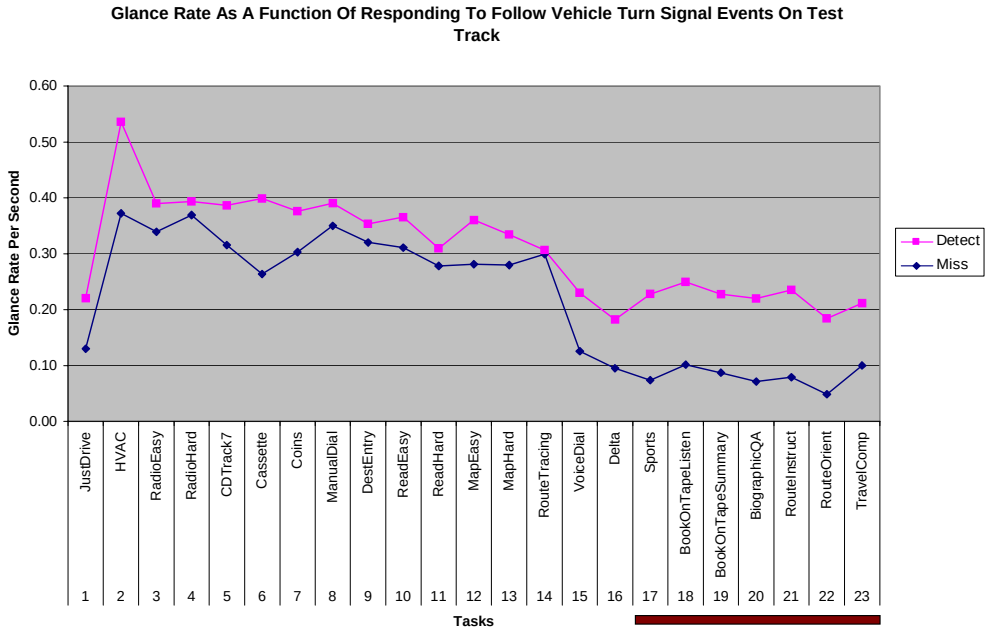


Figure 3-35. Task by Event Response Interaction for FVTS Events on the Glance Rate Metric

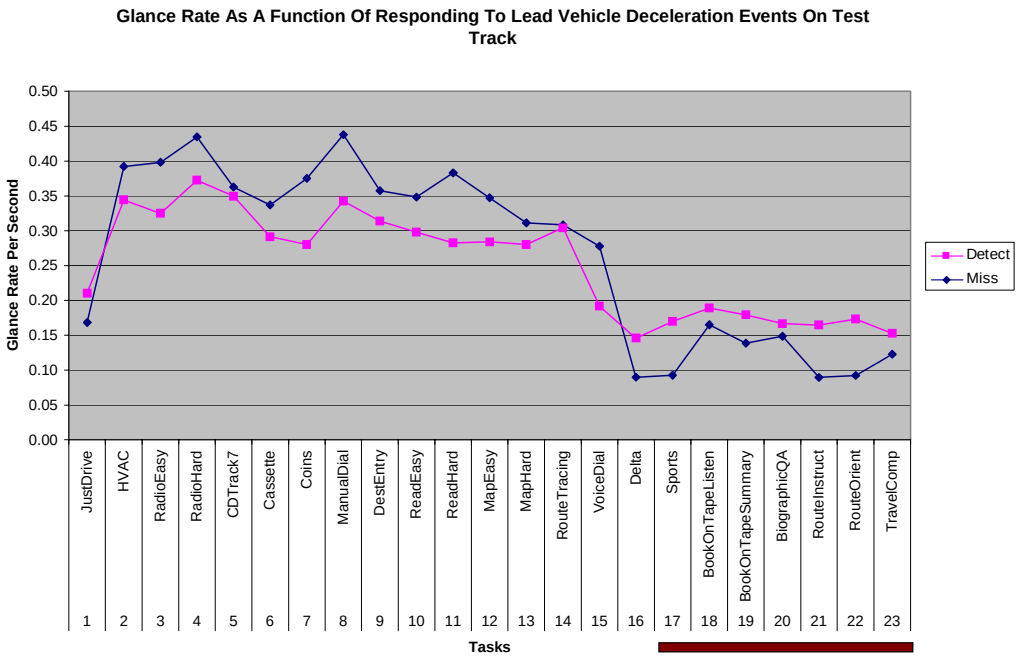


Figure 3-36. Task by Event Response Interaction for LVD Events on the Glance Rate Metric

3.4.2.6 Percent/Proportion of Task Duration Spent at Each Location

Figure 3-37, Figure 3-38, and Figure 3-39 depict the metric of Proportion of Task Duration Spent Looking at various location types (roadway, SA, and task-related) that was discussed earlier but are expressed in these figures as a percent and related to event detection. These figures are plotted in terms of drivers' response to CHMSL events (for illustration). Figure 3-37 shows that when a CHMSL was detected and responded to, there was an increase in the percent of task time spent viewing the roadway of 13 percent (67% versus 54% when the CHMSL was not responded to), 5 percent more was spent on SA locations, and 6 percent less on task-related locations. Figure 3-38 shows the results for FVTS events. The increase in percent of task time spent viewing the road was smaller, 4 percent (67% versus 63% when the FVTS was not responded to), the increase in time spent on SA locations was 7 percent, and the decrease in time spent on task-related locations was 4 percent. Figure 3-39 shows the results for LVD events. Time spent viewing the road increased by 8 percent, SA by 2 percent, and task-related areas decreased by 2 percent.

In general, the data indicate the following. Relative to "Miss" (no detect) trials, detection trials tended to be associated with a higher percentage of task duration spent looking at the road and situation awareness locations. Relative to "Miss" trials, detection trials were also associated with a smaller percentage of task duration spent looking at task-related locations.

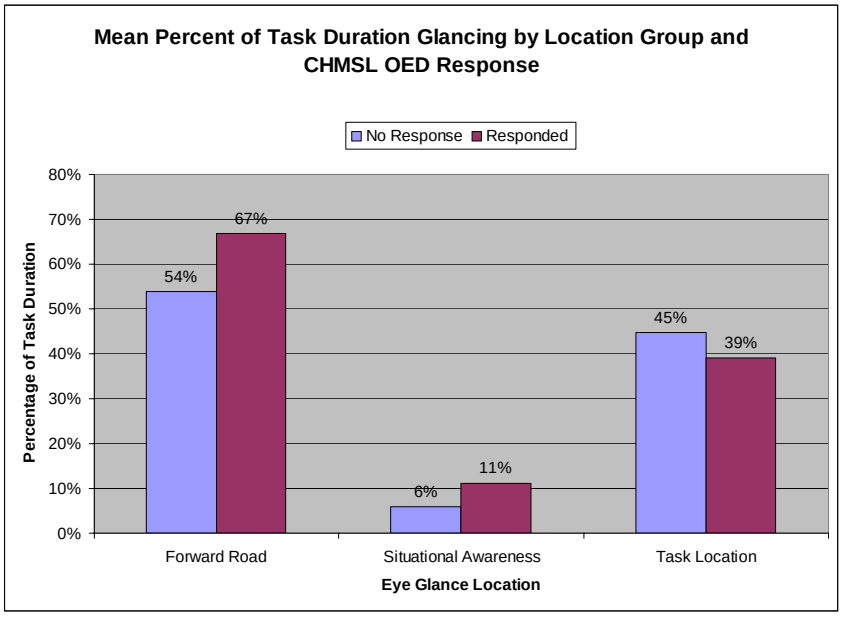


Figure 3-37. Mean Percent of Task Duration Spent Glancing at Roadway (Track), SA, and Task Locations
(as a function of whether or not CHMSLs were detected and responded to)

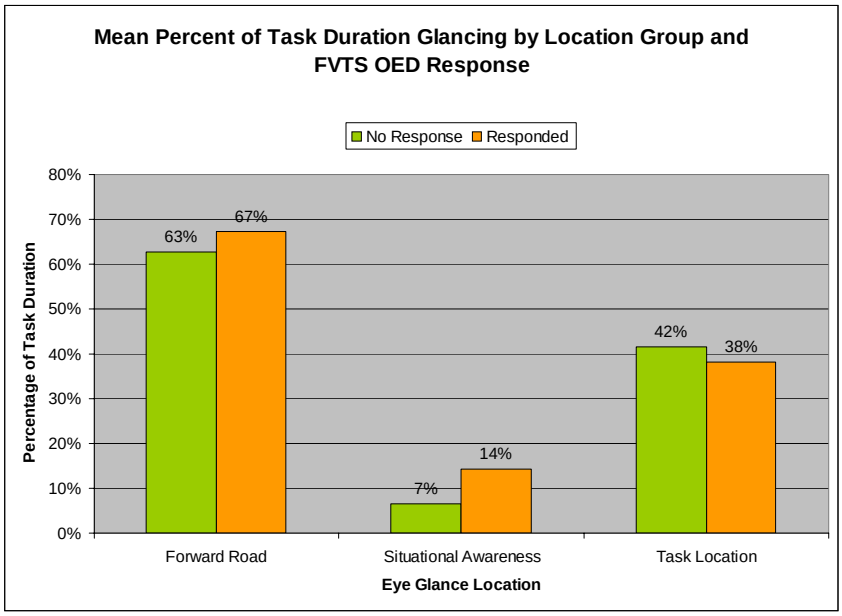


Figure 3-38. Mean Percent of Task Duration Spent Glancing at Roadway (Track), SA, and Task-Related Locations
(as a function of whether or not FVTS events were detected and responded to)

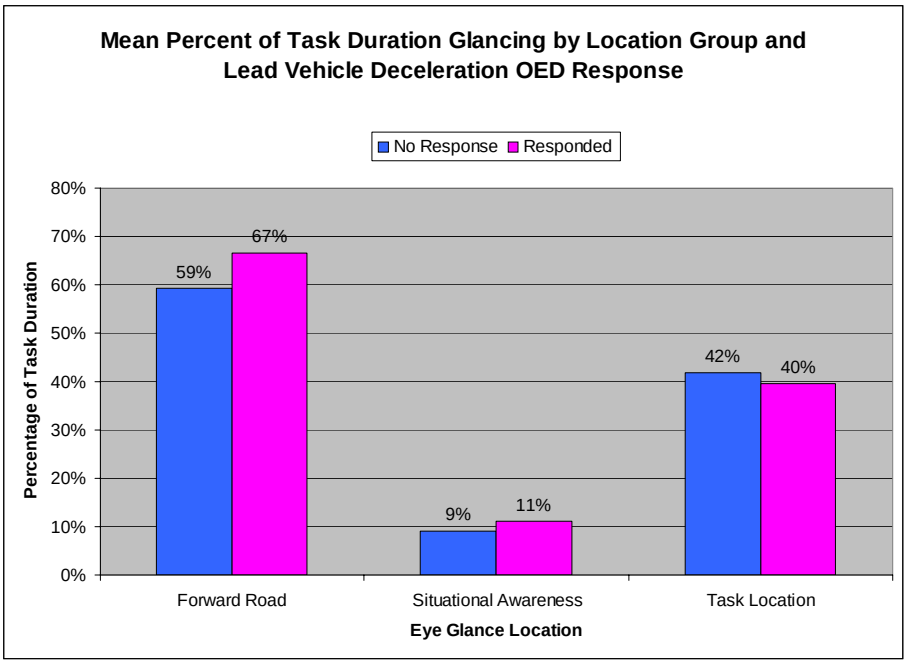


Figure 3-39. Mean percent of Task Duration Spent Glancing at Roadway (Track), SA, and Task-Related Locations
(as a function of whether or not a LVD event was detected and responded to)

3.4.2.7 Summary of Event Detection and Glance Pattern Relationships

In summary, formal analyses on the effects of event detection on glance patterns suggested that when events were detected and responded to by drivers, scan patterns subsequently changed in a way that was adaptive to the specific type of event to which the driver had responded. This finding has theoretical, methodological, and practical importance, and deserves considerable study and confirmation in future work.

Three eyeglance measures were broken out between "Detect" and "Miss" (no detect) trials and examined task-by-task for CHMSL, FVTS, and LVD stimulus events. The three selected measures were: a) Mean Single-Glance Duration to any Location, b) Maximum Glance Duration to any Location, and c) Glance Rate to any Locations. There were substantial differences between "Detect" and "Miss" trials on all three measures and all OED stimulus events for the auditory-vocal and Just Drive tasks (Figure 3-30 through Figure 3-36). There were no substantial differences for visual-manual tasks in mean single-glance duration or maximum single-glance duration for any OED stimulus events. There were no substantial differences for visual-manual tasks in glance rates associated with either CHMSL or FVTS stimulus events. There was a substantial difference among visual-manual tasks with glance rates and LVD events. The pattern of results indicated lower glance rates associated with "Detect" LVD events for visual-manual tasks. This may have been due to longer glances to the road in response to a lead vehicle deceleration. Longer glances to the road might have been needed to pick up looming or other visual cues of lead vehicle deceleration. Glance rates would have been reduced because the number of glances per unit time was reduced. This hypothesis would be strengthened by evidence of increased mean single-glance times to the road for "Detect" trials versus "Miss" (non-detect)

trials. This merits further investigation, which should include evaluation of other eyeglance measures such as task-related number of glances and task-related total glance durations.

Theoretically, the implications of these findings are that event detection may serve as an attentional interrupt for auditory-vocal tasks and the task of just driving, resulting in more active scanning of the road and mirrors for SA. The same phenomenon may hold only with certain types of OED stimulus events for visual-manual tasks. For visual-manual tasks, this appeared to occur only for LVD events. Hypotheses about the effects of events on the deployment of attention during driving need to be developed and confirmed in further work, particularly work that is done in a more naturalistic setting to see if event detection effects that were observed here were due somehow to the experimental paradigm or conditions used, or whether they will generalize to naturalistic driving.

From the point of view of methodology and practicality, there are several implications of event detection effects, if confirmed in further research, for measuring glance behavior in evaluations of advanced information systems. Eyeglance behavior collected during auditory-vocal tasks or Just Driving may need to involve trials with and trials without OED stimulus events. When evaluating the visual demand of tasks in an advanced information system or an in-vehicle device, it may be important that multiple test trials be conducted—some with and some without event detection. Under certain conditions indicated earlier, the trials used to evaluate the visual demand of a task should not include events-to-be-detected. Because the presence of event-to-be-detected can change durations and numbers of glances, depending on the type of event that is presented, these events can spuriously alter the visual-demand assessment results if they are included in trials used to assess visual demand. Ideally, an assessment would be done in a context in which drivers sometimes received events during tasks, and sometimes did not. The drivers would not know on which trials events would occur, and so would have to be monitoring for events on each trial. However, on the test trials actually used to assess visual demand for a task, no event would be presented. These trials would yield clean measurements of glance behavior, free from the influence of co-occurring events.

These results are both compelling and plausible. Further research is needed to replicate the findings. In addition, a fine-grained time series analysis is needed to relate eyeglance behavior to specific events. It is possible that the pattern of eyeglance behaviors generally do not reflect the impact of OED stimulus detection. Rather, it is the OED stimulus detection that reflects the general pattern of eyeglance behavior. This seems unlikely for the LVD detections during visual-manual tasks, but it is possible for the other results. A time-series analysis of each participant's eyeglances for each trial of a task's duration is needed to determine the prevalence of patterns to support either explanation.

3.4.3 Analyses of Reliability and Predictive Validity for Glance Metrics

In evaluating the properties of the measures taken on the test track, analyses of reliability and predictive validity for the eyeglance measures were undertaken and were consistent with those done on other categories of measurement.

3.4.3.1 Overall Level—Split-half Reliability

To examine the reliability of eyeglance measures, the sample of data collected on the test track was split in half and correlations between the split halves were computed, following the methods previously described. However, it should be noted that eye data from the track came from only 18 research participants. That is, up to 9 participants' data were used to calculate a task's summary statistic for use in the analysis. This meant that the split halves were well balanced for the eye data analyses, but not perfectly balanced by age and gender as they are for most other subsets of

the data. Nonetheless, there was a desire for all split-half analyses to be similarly implemented. The split-half correlations were done across the full set of tasks performed on the test track.

The outcomes of the correlations between the split halves are shown in Table 3-5. (The full set of eyeglance metrics and their abbreviated names were presented in Table 3-2.) The extended variable names in the leftmost column of Table 3-5 are shown for those variables that proved reliable across both test track and road venues (with correlations greater than ± 0.707). The correlations are highlighted in green in the table. Generally, the eyeglance measures that met this criterion for reliability fell into a small number of groups. Metrics related to the following categories were reliable:

- Number of glances:
  - to road locations;
  - to situation awareness locations;
  - to task-related areas; and
  - to Total/All, and To Not Road (combines everything other than road).
- Durations of glances:
  - mean (for most locations: Road, Situation Awareness, Task (was borderline), and Not Road);
  - median (for some locations: Road, Situation Awareness);
  - standard deviation (for some locations: Road, Task, Total/All, Not Road); and
  - max (for only certain location types: Road, Task, Total/All).
- Accumulations of durations for certain location types:
  - total glance time to road location;
  - total glance time to situation awareness location; and
  - total glance time to task-related areas.
- Percents (or proportions) of task time spent looking at a location type:
  - to road locations;
  - to situation awareness areas (borderline); and
  - to task-related areas.
- Rates of glances per second:
  - overall (Total/All), Road, Task, Not Road.

Table 3-5. Split-Half Correlations for Test Track Data on Eyeglance Measures

Eye Glance Metric	TRACK DATA Split-Half Reliability, Pearson r	Eye Glance Metrics Repeatable In Test Track Venue
MeanTskglncs	0.852	**Total Glances to All Locations During Task**
MeanTaskdur	0.996	**Mean Task Duration Derived From Eye Data**
MeanmeanTdur	0.681	**Mean of Mean Duration Of All Glances (To All Locations) During Task**
MeanmedTdur	0.518	
MeansdTdur	0.829	**Mean Stand. Deviation of Glance Durations To All Locations During Task**
MeanTglsprs	0.930	**Mean Rate of Glances Per Second During Task (includes glances to all locations)**
MeanglncesRD	0.852	**Mean Number of Glances To Road**
MeanduratRD	0.996	**Mean Total Glance TimeTo Road During Task (Summed Across Glances)**
MeanmeanRDdr	0.811	**Mean of Mean Glance Durations To Road**
MeanmedRDdur	0.788	**Mean Of Median Glance Durations To Road**
MeansdRDdur	0.839	**Mean Stand. Deviation of Glance Durations To Road**
MeangrateRD	0.917	**Mean Glance Rate Per Second To Road (During Task)**
MeanpctdurRD	0.983	**Mean Percent of Task Duration Spent Looking At Road**
MeanglncesSA	0.951	**Mean Number of Glances To Sit Awareness Locations (Mirrors & Speedo)**
MeanduratSA	0.976	**Mean Total Glance Time To Sit Awareness Locations (Summed Across Glances)**
MeanmeanSAdr	0.828	**Mean of Mean Glance Durations To Sit Awareness Locations**
MeanmedSAdur	0.781	**Mean of Median Glance Durations To Sit Awareness Locations**
MeansdSAdur	0.548	
MeangrateSA	0.245	
MeanpctdurSA	0.675	**Mean Percent Duration of Task Spent Looking At Sit Awareness Locations**
MeanglncesTR	0.983	**Mean Number of Glances To Task-Related Areas**
MeanduratTR	0.977	**Mean Total Glance Time to Task-Related Areas (Summed Across Glances)**
MeanmeanTRdr	0.677	**Mean of Mean Duration of Glances To Task-Related Areas**
MeanmedTRdur	0.599	
MeansdTRdur	0.784	**Mean Stand. Deviation of Glance Durations To Task-Related Areas**
MeangrateTR	0.932	**Mean Glance Rate Per Second To Task-Related Areas**
MeanpctdurTR	0.939	**Mean Percent Of Task Duration Spent Looking At Task Locations**
MeanglncesNA	0.147	
MeanduratNA	0.661	
MeanmeanNAdr	0.497	
MeanmedNAdur	0.455	
MeansdNAdur	-0.112	
MeangrateNA	0.904	**Mean Glance Rate Per Second During Task Spent Looking At N.A./Obstructed**
MeanpctdurNA	0.940	**Mean Percent Duration of Task Spent Looking At NA Locations**
MeanglncesMR	0.961	**Mean Number of Glances At Mirrors Alone**
MeanduratMR	0.984	**Mean Total Glance Time To Mirrors Alone (Summed Across Glances)**
MeanmeanMRdr	0.750	**Mean of Mean Glance Durations To Mirrors**
MeanmedMRdur	0.675	**Mean of Median Glance Durations To Mirrors**
MeansdMRdur	0.492	
MeangrateMR	0.077	
MeanpctdurMR	0.558	
MeanglncesNR	0.851	**Mean Number of Glances To All Areas Classified As "NOT ROAD"**
MeanduratNR	0.919	**Mean Total Glance Time To All "NOT ROAD" Areas (summed across glances)**
MeanmeanNRdr	0.728	**Mean of Mean Duration Of Glances To "NOT ROAD" Areas**
MeanmedNRdur	0.633	
MeansdNRdur	0.896	**Mean Stand. Deviation of Glance Durations To "NOT ROAD" Areas**
MeangrateNR	0.941	**Mean Glance Rate Per Second To "NOT ROAD" Areas**
MeanpctdurNR	0.980	**Mean Percent Time During Task Spent Viewing "NOT ROAD" Areas**
MinTdur	-0.145	
MinRDdur	-0.037	
MinSAdur	0.154	
MinTRdur	0.542	
MinNAdur	0.262	
MinMRdur	0.004	
MinNRdur	0.178	
MaxTdur	0.732	**Maximum Duration of Glances To All Locations During Task**
MaxRDdur	0.777	**Maximum Duration of Glances To The Road**
MaxSAdur	0.430	
MaxTRdur	0.892	**Maximum Duration of Glances To The Task**
MaxNAdur	0.393	
MaxMRdur	0.483	
MaxNRdur	0.367	

Items in green have "r" values >.707 (original cutoff for repeatability)

Items in blue have "r" values >0.665 and p<0.05

3.4.3.2 Predictive Validity – Correlations between Track Eyeglance Data and Track Driving and OED Performance Metrics

Measures of eyeglance behavior are considered fundamental to driving performance. It is, therefore, not necessary to establish whether they have predictive validity (or whether other driving performance measures can be predicted from eyeglance measures). Nonetheless, it is informative to explore the relationships that exist between eyeglance metrics and other driving performance measures. Such analyses provide insights that may help formulate multivariate analyses and may assist in pulling together a coherent picture of how task workload affects driving performance from this set of individual measurements.

Correlations for Full Set of Tasks (Between Eyeglance Data and Performance Data)

Table 3-6 presents the correlations between the eyeglance metrics and the reliable driving performance metrics across the full task set (auditory-vocal, visual-manual, and mixed-mode tasks) for the test track. This is important to keep in mind. For example, visual-manual tasks are associated with a back-and-forth glance pattern between road and task. Thus, more glances to the road are generally accompanied by more glances to the task as well. Auditory-vocal tasks do not have this property. For these reasons, the results discussed in this section will be broken out by task type.

Median standard deviation of lane position (SDLP) correlated only with mean number of glances made to any location throughout a task as a whole, the mean number of glances made to the road, and the mean number of glances made to the “not road” location. These are rather difficult to interpret in the absence of other significant relationships. As the number of glances to the road increased, the standard deviation of lane position increased. These relationships were surprising and seemed counterintuitive, though they may be at least partially due to task duration effects. Generally, shorter durations limit the magnitude of lane position variability, speed variability, range variability, and so on. This arises from the physical properties of the vehicle. Longer durations allow more time for such variations to increase in magnitude. A plausible interpretation for correlations between SDLP and glances to the road is that longer duration tasks were associated with laxer lanekeeping, either due to workload effects or the continuous effort required for crisp vehicle control over longer periods. The longer the task’s duration, the more glances and the higher the SDLPs that are possible. The expectation was originally that the more times the driver glanced at the road, the better lanekeeping would have been and the smaller SDLP would have been. But this was not the case, possibly for the explanation given above.

Another hypothesis might be that these relationships were indicative of a state in which some attention had been shifted to something other than lanekeeping, even though the number of glances to the forward road was high. One possibility is that attention was shifted in part to event monitoring, as opposed to being fully allocated to lanekeeping or the in-vehicle device activity. The correlation between number of task-related glances and SDLP did not by itself meet the criterion for highlighting in the table. However, looking at task-related areas for the in-vehicle activity may also have played some role in increasing variance in SDLP that is reflected in the number of glances to all locations and the number to the “not road” category, both of which were highly correlated with SDLP. When monitoring for events, drivers may have been glancing between mirrors and road. Glances to the mirrors would have increased the number of the glances to the road, and may have led to increases in SDLP (in the case where there were small inadvertent steering inputs when the eyes were on the mirrors, or in cases where the driver’s eyes were off the road more across multiple mirror glances and transitions, and some lane

position variation went unnoticed momentarily). However, this hypothesis requires further analysis and study.

Highlighted in a softer shade of green for Percent Trials with a Cross of the Lane Line are two correlations, both with eyeglance metrics related to task-related glances. These are Mean Number of Glances to Task-Related Areas (MeanGlancesTR) and Mean Total Glance Time to Task-Related Areas (MeanDuratTR). These are noteworthy, insofar as glances to the in-vehicle task were associated to some extent with excursions from the lane.

The Median Speed Difference variable correlated very consistently with the mean number of glances to the road and their durations, and with the mean number of glances to the mirrors and their durations, as well as with the overall number and duration of glances to all locations during a task. (Glances associated with mirrors were measured in two ways, one using the SA location type, which included both glances to mirrors and speedometer and one with the MR location type, which included glances to mirrors only.) Of the two, the correlations with mean glances and durations to the mirrors were slightly stronger. These relationships may also be interpreted to indicate that drivers were monitoring for events, using glances to the road and mirrors. The more glances and the longer the glances, however, the larger the speed difference during the task. On the face of it, this seems like a somewhat counterintuitive result. However, Speed Difference was also influenced by task duration. There was more opportunity for Speed Difference to grow as task duration increased. The paradoxical results of OED for shorter tasks suggests this as a credible hypothesis. As suggested previously, it is also possible that these correlations may hint at a state of monitoring for events, that is multitasked with monitoring speed and lanekeeping, i.e., with driving, but such a hypothesis requires further analysis and study for verification.

An alternative hypothesis regarding why SDLP and Speed Difference increased with more glances to the road and mirrors is that perhaps drivers became more visually attentive to the road (and showed more variable steering and speed) when they subjectively felt as if an in-vehicle activity imposed a high workload. This hypothesis was tested with a correlation between Median Operator Workload Level (OWL) ratings for each task and Median SDLP, r = -0.193, as well as between Median OWL ratings and Median Speed Diff, r = -0131. In both cases, the correlations were low and in a direction opposite to that predicted by a hypothesis based on perceived workload level. Such a hypothesis cannot explain the correlations observed between the eye metrics and the performance data.

The measures of driver responsiveness to events (Percent LVD Miss Rate, Percent CHMSL Miss Rate, and Percent FVTS Miss Rate) also correlated in a very consistent way with the eyeglance measures, though the correlations were the strongest for Percent CHMSL Miss Rate. The highest positive correlations were with glance rate per second for the total task (all glances included, regardless of location), with glance rate to the road, and with glance rate to task-related areas. (Note that these measures are correlated and do not provide independent information. For example, glance rates to all locations include glance rates to the road and glance rates to the tasks). These relationships indicated that the higher the glance rate, the higher the miss rate. As seen in the graphs depicting the Task by Location Type interaction, high-glance rates to the task are associated with high-glance rates to the road—the pattern of looking back and forth between task and road. This may again be an instance of a relationship specific to a subset of tasks dominating the full task set in the computation of overall correlations (namely, the visual-manual subset of tasks).

Strong negative correlations emerged for the various metrics associated with durations of glances to the road and to the mirror/situation awareness locations (mean, median, standard deviation, and accumulated duration across task). These negative correlations indicate that the shorter the glances are to the road and/or mirror locations, the higher the miss rates are for CHMSLs, FVTSs, and LVD. Short glances to the road and mirrors may be associated with very high glance rates as might have occurred, for example, when the driver was performing a visual-manual task, and monitoring for events, as well as monitoring lane position and headway. In this situation, many locations were being scanned, and the rate of glancing was high, with durations that were shorter, and more time spent in transition. Both shorter time on the roadway and on mirror locations may have increased miss rates. Higher glance rates may also be associated with more transition times that may themselves lead to increased miss rates (since vision is suppressed when the eye is moving). These relationships would benefit from further study at a time-history level.

Table 3-6. Correlations between Eyeglance Metrics and Reliable Driving Performance Metrics Across the Full Task Set for the Test Track

Note: Positive correlations over +0.707 are highlighted in green; negative correlations less than -0.707 are highlighted in yellow. Weaker correlations are highlighted in softer hues of these colors.

Test Track Correlations For Full Set Of Tasks						
	Median SDLP	*Median Speed Diff*	*%Cross Trials*	*%LVD Miss Rate*	*%CHMSL Miss Rate*	*%FVTS Miss Rate*
MeanTskglncs	0.778	0.840	0.406	-0.543	-0.446	-0.330
MeanTaskdur	0.651	0.841	-0.018	-0.694	-0.798	-0.583
MeanmeanTdur	0.202	0.531	-0.353	-0.584	-0.739	-0.579
MeansdTdur	0.281	0.628	-0.344	-0.653	-0.806	-0.605
MeanTglsprs	-0.308	-0.636	0.359	0.667	0.809	0.618
MeanglncesRD	0.781	0.855	0.376	-0.569	-0.485	-0.369
MeanduratRD	0.528	0.739	-0.167	-0.673	-0.837	-0.625
MeanmeanRDdr	0.241	0.578	-0.351	-0.618	-0.766	-0.611
MeanmedRDdur	0.205	0.549	-0.369	-0.593	-0.738	-0.592
MeansdRDdur	0.275	0.618	-0.312	-0.654	-0.812	-0.603
MeangrateRD	-0.336	-0.661	0.344	0.680	0.816	0.619
MeanpctdurRD	0.223	0.550	-0.343	-0.652	-0.808	-0.638
MeanglncesSA	0.557	0.733	-0.152	-0.684	-0.794	-0.729
MeanduratSA	0.528	0.699	-0.173	-0.668	-0.783	-0.755
MeanmeanSAdr	0.322	0.534	-0.312	-0.648	-0.785	-0.833
MeanmedSAdur	0.292	0.504	-0.334	-0.634	-0.756	-0.827
MeanpctdurSA	0.134	0.293	-0.361	-0.496	-0.639	-0.823
MeanglncesTR	0.507	0.423	0.685	-0.097	0.202	0.253
MeanduratTR	0.492	0.407	0.684	-0.071	0.222	0.271
MeanmeanTRdr	0.108	-0.465	0.462	0.584	0.681	0.670
MeangrateTR	-0.099	-0.481	0.436	0.533	0.740	0.566
MeanpctdurTR	-0.069	-0.443	0.466	0.551	0.752	0.581
MeangrateNA	-0.589	-0.876	-0.078	0.877	0.786	0.628
MeanpctdurNA	-0.195	-0.390	-0.219	0.440	0.492	0.378
MeanglncesMR	0.575	0.757	-0.135	-0.694	-0.795	-0.716
MeanduratMR	0.546	0.720	-0.160	-0.678	-0.786	-0.747
MeanmeanMRdr	0.404	0.601	-0.223	-0.672	-0.751	-0.849
MeanmedMRdur	0.382	0.584	-0.228	-0.667	-0.714	-0.842
MeanglncesNR	0.767	0.822	0.436	-0.512	-0.391	-0.291
MeanduratNR	0.631	0.605	0.574	-0.257	-0.037	0.016
MeanmeanNRdr	-0.118	-0.277	0.029	0.409	0.495	0.448
MeansdNRdur	-0.024	-0.223	0.004	0.243	0.411	0.360
MeangrateNR	-0.311	-0.632	0.353	0.665	0.819	0.620
MeanpctdurNR	-0.234	-0.549	0.333	0.645	0.813	0.635
MaxTdur	0.422	0.646	-0.337	-0.653	-0.825	-0.668
MaxRDdur	0.402	0.636	-0.288	-0.644	-0.830	-0.687
MaxTRdur	0.194	-0.170	0.680	0.300	0.497	0.518
		Highlights + correlations >.707				Highlights - correls < -.707
		Highlights + correlations >.665, p<.05				Highlights - correls < -..665

Correlations Across Subsets by Task Type

To clarify the interpretation of the correlations done across the entire task set, additional correlations were done on smaller subsets of tasks. Specifically, correlations were separately done on the visual-manual tasks, the auditory-vocal tasks, and on the mixed-mode tasks (plus Just Drive). Though Just Drive is quite different from both task types,

and is not a mixed-mode task, it was grouped with them to enable examination of the remaining variance after partialing out the visual-manual and auditory-vocal tasks in this series of analysis.

Visual-Manual Tasks

For Median SDLP, as expected, the correlations remained with mean number of glances made to any location throughout a task as a whole and to mean number of glances made to the road. See Table 3-7. However, additional correlations emerged, among them correlations (though weaker than the +0.707 cutoff) with two measures of task-related glances (Mean Number of Glances to Task-Related areas, and Total Glance Time to the Task). This lends some support to the notion that glances to the task are playing a role. However, there were other correlations that emerged as well. These were between number of glances to the road, SA, and MR locations, as well as the durations of those glances. Together these correlations clarify the picture. The more glances that are made to acquire/maintain/update situation awareness and detect events (by looking at the road and the mirrors), the larger SDLP becomes. This is not intuitive, though several hypotheses can be generated to explain it. The correlation between average glance counts to the road and average glance counts to task-related locations, across the 13 tasks, was almost perfect ($r = 0.99$ $p < 0.5$). Instead, it may be that other factors are involved, such as checking for events while driving and doing an in-vehicle activity.

Some correlations emerged between glance measures and percent trials with a cross of the lane line that were not present for the overall data set. This suggests that the relationships between glance measures and lane departure metrics (often reported previously in U.S. literature) were present for the visual-manual subset of tasks in this study too. The metric of Percent Cross Trials was related to total duration spent looking at the road (highlighted) and less strongly (not highlighted, but worth mentioning) to mirrors/situation awareness areas, and to some measures of task-related glances.

Surprisingly, the relationships between eye behavior and responsiveness to events have all but disappeared in the data for visual-manual tasks only. This is consistent with the findings of Young and Angell (2003), that event-detection for visual-manual tasks was only weakly predicted by traditional eyeglance measures. In the present study, which extended the set of eyeglance metrics beyond those typically obtained, however, there were two exceptions. For Percent LVD Miss Rate, when the Mean Glance Rate to the NA area (obstructed or not able to be scored) was high, the Percent LVD Miss Rate was high. This occurred on tasks requiring paper materials to be used, since drivers sometimes held them in front of their faces and obstructed their own view of the forward roadway (leading them to miss seeing LVD events). The other relationship between eye behavior and event detection for visual-manual tasks was between duration of looks to the Mirror and SA locations (i.e., mirrors and speedometer). The shorter the glances to the mirrors, the higher the FVTS miss rate, which makes sense, since FVTS events were detected in the left outside mirror.

Table 3-7. Correlations between Eyeglance Metrics and Other Driving Performance Metrics for Visual-Manual Tasks Only

Correlations for Test Track: Visual-Manual Tasks Only From Test Track Data						
	Median SDLP	*Median Speed Diff*	*%Cross Trials*	*%LVD Miss Rate*	*%CHMSL Miss Rate*	*%FVTS Miss Rate*
MeanTskglncs	0.706	0.914	0.633	-0.486	-0.173	-0.005
MeanTaskdur	0.737	0.927	0.638	-0.493	-0.187	-0.008
MeanmeanTdur	0.072	0.000	-0.392	-0.094	-0.203	-0.164
MeansdTdur	0.111	0.071	-0.398	-0.216	-0.275	-0.217
MeanTglsprs	-0.254	-0.028	0.119	-0.052	0.030	0.136
MeanglncesRD	0.708	0.915	0.636	-0.489	-0.172	-0.005
MeanduratRD	0.781	0.918	0.702	-0.531	-0.226	-0.085
MeanmeanRDdr	-0.047	-0.246	-0.115	0.057	-0.161	-0.236
MeanmedRDdur	-0.076	-0.258	-0.144	0.101	-0.100	-0.206
MeansdRDdur	-0.026	-0.222	-0.134	0.006	-0.218	-0.273
MeangrateRD	-0.375	-0.184	0.049	0.055	0.091	0.180
MeanpctdurRD	-0.155	-0.339	-0.030	0.065	-0.167	-0.231
MeanglncesSA	0.758	0.947	0.593	-0.608	-0.216	-0.171
MeanduratSA	0.775	0.948	0.614	-0.618	-0.243	-0.224
MeanmeanSAdr	-0.046	-0.143	-0.142	-0.162	-0.324	-0.699
MeanmedSAdur	-0.087	-0.207	-0.208	-0.128	-0.245	-0.671
MeanpctdurSA	-0.417	-0.448	-0.310	0.044	-0.264	-0.499
MeanglncesTR	0.691	0.899	0.636	-0.459	-0.165	0.027
MeanduratTR	0.683	0.890	0.638	-0.434	-0.147	0.054
MeanmeanTRdr	0.022	0.036	0.323	0.308	0.386	0.341
MeangrateTR	0.165	0.380	0.437	-0.232	0.000	0.322
MeanpctdurTR	0.204	0.362	0.441	-0.037	0.193	0.350
MeangrateNA	-0.708	-0.819	-0.660	0.746	0.464	0.183
MeanpctdurNA	-0.049	-0.077	-0.542	0.094	0.149	-0.010
MeanglncesMR	0.755	0.945	0.588	-0.605	-0.217	-0.177
MeanduratMR	0.775	0.947	0.610	-0.610	-0.230	-0.225
MeanmeanMRdr	0.256	0.126	0.099	-0.265	-0.214	-0.701
MeanmedMRdur	0.273	0.130	0.072	-0.279	-0.144	-0.685
MeanglncesNR	0.702	0.910	0.628	-0.482	-0.164	-0.005
MeanduratNR	0.693	0.912	0.587	-0.474	-0.163	0.025
MeanmeanNRdr	0.106	0.219	-0.305	-0.148	-0.049	0.017
MeansdNRdur	0.185	0.242	-0.297	-0.281	-0.193	-0.086
MeangrateNR	-0.222	-0.003	0.126	-0.061	0.025	0.153
MeanpctdurNR	0.151	0.339	0.013	-0.083	0.150	0.232
MaxTdur	0.148	0.166	-0.390	-0.215	-0.204	-0.108
MaxRDdur	0.038	-0.064	-0.077	-0.015	-0.299	-0.471
MaxTRdur	0.483	0.514	0.677	-0.304	-0.171	0.168
		Highlights + correlations > 0.707			Highlights - correlations <-.707	
		Highlights + correlations >0 .665 but <.707			Highlights - correlations <-0.665 but >-0.707	

Auditory–Vocal Tasks

Correlations for the auditory-vocal tasks, Table 3-8, showed some relationships similar to visual-manual tasks with regard to SDLP. However, task-related glances cannot be analyzed meaningfully for auditory-vocal tasks. A low incidence of task-related glances for auditory-vocal tasks was noted in Section 3.4.1.4. The proportions of task-related glances approached 0.00 for these tasks and this implies zero or near-zero number of task-related glances, as defined in this project. This point is made again later as well. This precludes any comment on any relationship that is dependent upon task-related glances. The relationship between number of glances and their durations to road and mirrors/situation awareness areas are stronger than in the full set of tasks or in the visual-manual subset. This may be expected since there are generally no other glance locations for the driver to look at. These relationships suggest that the more glances to the road and mirrors, and the longer they are, the larger SDLP is. This result is again counterintuitive. It may be suggestive of an underlying "satisficing" process in which the driver feels more aware of the road and relaxes lanekeeping tolerances somewhat, especially as the task lengthens (auditory-vocal tasks, with one exception, were ~2 minutes in length). Alternatively, longer duration tasks may provide less driver discretion and so may be

associated with laxer vehicle control because of workload effects (for auditory-vocal tasks) lost-in-thought distraction effects (for Just Drive), or the continued effort required for crisp vehicle control. This is unclear and deserves further investigation.

The relationships between eye behavior and responsiveness to events in the correlations were strongest for the FVTS events. Most interesting in this regard is the strong positive correlation between the Percent of Task Time Spent Looking at the Road and Percent FVTS Miss Rate, substantiating the effects noted in the univariate results previously discussed. During auditory-vocal tasks, as drivers concentrated gaze for prolonged periods on the forward roadway, scanning of the outside mirrors dropped somewhat, and miss rates went up for the FVTS events. The negative correlations between number of glances to the road and miss rates also is consistent with this (as the number of glances to the road go down, more steady gazing is occurring, and miss rates go up). Other correlations, highlighted in yellow, include number of glances to the mirrors and SA locations (as there are fewer of them, miss rates for FVTS events increased).

One note of caution, there are some correlations highlighted in the table for task-related eye metrics (MeanduratTR, MeanmeanTRdr, MeangrateTR). As mentioned, there were a few, infrequent glances upward made during a subset of auditory-vocal tasks. For these few tasks, there were on average, only 1 or 2 per task. The correlations that emerged in the analysis are thus based on these few glances, which occurred repeatedly within the small sample of subjects for a few of these tasks, but were nonetheless few in number.

Table 3-8. Correlations Between Eyeglance Metrics and Driving Performance Metrics for Auditory-Vocal Tasks Only

Correlations For Test Track: Auditory-Vocal Tasks Only From Test Track Data						
	Median SDLP	*Median Speed Diff*	*% Cross Trials*	*% LVD Miss Rate*	*% CHMSL Miss Rate*	*% FVTS Miss Rate*
MeanTskglncs	0.915	0.779	0.265	-0.400	-0.205	-0.842
MeanTaskdur	0.961	0.902	0.338	-0.406	-0.356	-0.723
MeanmeanTdur	0.036	0.093	0.034	-0.115	-0.648	-0.276
MeanTglsprs	0.192	-0.155	-0.173	-0.150	0.426	-0.743
MeanglncesRD	0.918	0.780	0.268	-0.405	-0.211	-0.840
MeanduratRD	0.963	0.913	0.343	-0.406	-0.382	-0.705
MeanmeanRDdr	0.186	0.323	0.120	-0.108	-0.674	-0.340
MeanmedRDdur	0.002	0.165	-0.146	0.083	-0.467	-0.304
MeansdRDdur	0.283	0.467	0.732	-0.324	-0.631	-0.106
MeangrateRD	-0.027	-0.376	-0.256	-0.079	0.531	-0.593
MeanpctdurRD	-0.777	-0.568	-0.198	0.340	-0.074	0.861
MeanglncesSA	0.917	0.790	0.272	-0.433	-0.203	-0.850
MeanduratSA	0.917	0.792	0.312	-0.448	-0.183	-0.839
MeanmeanSAdr	0.854	0.804	0.614	-0.532	-0.601	-0.595
MeanmedSAdur	0.808	0.763	0.675	-0.562	-0.647	-0.537
MeanpctdurSA	0.695	0.438	0.229	-0.493	0.123	-0.916
MeanglncesTR	0.489	0.175	0.443	-0.466	-0.383	0.258
MeanduratTR	0.264	-0.041	0.225	-0.342	-0.344	0.472
MeanmeanTRdr	-0.672	-0.767	-0.666	0.290	-0.006	0.940
MeangrateTR	-0.538	-0.764	-0.415	-0.031	0.143	0.802
MeanpctdurTR	-0.497	-0.726	-0.393	-0.033	0.094	0.810
MeangrateNA	-0.835	-0.747	-0.578	0.679	0.641	0.680
MeanpctdurNA	-0.269	0.133	-0.385	0.701	-0.210	0.415
MeanglncesMR	0.906	0.787	0.249	-0.416	-0.187	-0.861
MeanduratMR	0.910	0.790	0.290	-0.431	-0.162	-0.848
MeanmeanMRdr	0.798	0.786	0.673	-0.535	-0.592	-0.564
MeanmedMRdur	0.711	0.704	0.768	-0.595	-0.643	-0.476
MeanglncesNR	0.912	0.776	0.248	-0.389	-0.199	-0.847
MeanduratNR	0.924	0.815	0.277	-0.382	-0.196	-0.813
MeanmeanNRdr	0.172	0.547	-0.127	0.408	-0.458	0.097
MeanmedNRdur	0.097	0.446	-0.154	0.449	-0.486	0.090
MeansdNRdur	0.418	0.501	0.312	-0.373	0.200	-0.012
MeangrateNR	-0.034	-0.354	-0.332	-0.013	0.586	-0.600
MeanpctdurNR	0.587	0.344	-0.014	-0.214	0.344	-0.846
MaxTdur	0.799	0.638	0.349	-0.409	-0.530	-0.753
MaxRDdur	0.799	0.638	0.349	-0.409	-0.530	-0.753
MaxTRdur	0.295	-0.036	0.294	-0.426	-0.281	0.405
		Highlights + correlations >.707				Highlights - correls < -.707
		Highlights + correlations >.665, p<.05				Highlights - correls < -..665

Mixed-Mode Tasks with Just Drive

Correlations with the remaining subset of tasks, the mixed-mode tasks of Voice Dial and Delta Flight Information plus Just Drive (a very different task involving no additional in-vehicle tasking, but involving specifically the response to external events while driving), revealed some very interesting patterns (see Table 3-9).

Note: Eyeglance metrics related to glances made to task-related areas are omitted from this table, due to the fact that the Just Drive task had no such glances. Also, correlations of 1.0, -1.0, or 0.0 result from rounding from 7 digits to the 3 that are reported in the table.

There were multiple high correlations between glance measures and SDLP and Speed Difference metrics, consistent with those previously discussed.

Most interesting in this set of correlations, were the correlations emerging between event detection measures and the eye metrics. There were very strong relationships between glance rate-per-second for the task (all glances to all locations) and glance rate-per-second to the road, and the Percent CHMSL Miss Rate (the higher the rate of glancing, the higher the miss rate). As noted earlier, transitioning rapidly between locations may thus be associated with missing events. This finding merits further exploration since vision is suppressed while the eye is moving and more transitions may thus have a consequence on event detection. On the other hand, the OED event lasted considerably longer than a typical eyeglance transition time. The glance rate measure, as explained earlier in this chapter, may be less suitable or interpretable than some other measures. For example, the number of glances and duration of glances to the road and to the mirrors/situation awareness areas were negatively correlated with the Percent CHMSL Miss Rate (the more glances and the longer the glances, the lower the CHMSL miss rate). For the Percent FVTS Miss Rate, the longer a driver looked at the mirror (mean of mean SA duration), the lower the Percent FVTS Miss Rate. For Percent LVD Miss Rate, the higher the glance rate to the road, the lower the percent miss rate for LVDs. However, it was also the case that the longer the mirror glances, the lower the LVD miss rates. The correlations in this table suggest that the Just Drive task contributed in a prominent way to the effects of event detection that were present in the correlation analysis on the full task set (especially for CHMSL and LVD events). This may have been true to some extent also for effects of tasks on lateral/longitudinal vehicle control effects: the Just Drive task appears to have contributed strongly. This possibility, that there are pronounced effects of event detection on eyeglance behavior from the Just Drive task, is fascinating and perhaps important, since during Just Drive, event detection was the one other activity competing for driver attention; whereas during the other tasks, driver attention was split in more ways, in order to carry out an in-vehicle task as well. This may have impacted the overall level of attention allocated to each activity being concurrently performed, and may have resulted in the performance decrements noted in the analyses described in this chapter.

Table 3-9. Correlations Between Eyeglance Metrics and Driving Performance Metrics for Mixed-Mode Tasks and the Just Drive Task for Test Track (Based on Voice Dial, Delta Flight Information, and Just Drive Tasks)

Test Track Correlations with Combination Tasks and Just Drive						
	Median SDLP	*Median Speed Diff*	*% Cross Trials*	*% LVD Miss Rate*	*% CHMSL Miss Rate*	*% FVTS Miss Rate*
MeanTskglncs	0.997	-0.380	0.906	0.423	-1.000	0.130
MeanTaskdur	0.994	-0.200	0.969	0.585	-0.984	0.313
MeanmeanTdur	0.597	-0.947	0.276	-0.404	-0.652	-0.661
MeansdTdur	0.852	-0.760	0.612	-0.037	-0.886	-0.337
MeanTglsprs	-0.981	0.115	-0.987	-0.653	0.965	-0.394
MeanglncesRD	0.992	-0.426	0.883	0.377	-0.998	0.080
MeanduratRD	0.992	-0.181	0.974	0.601	-0.980	0.332
MeanmeanRDdr	0.729	-0.875	0.441	-0.237	-0.775	-0.519
MeanmedRDdur	0.671	-0.911	0.367	-0.315	-0.721	-0.587
MeansdRDdur	0.975	-0.510	0.834	0.287	-0.988	-0.015
MeangrateRD	-0.955	0.009	-0.998	-0.729	0.932	-0.489
MeanpctdurRD	0.999	-0.256	0.953	0.537	-0.993	0.258
MeanglncesSA	0.855	-0.756	0.617	-0.031	-0.889	-0.331
MeanduratSA	0.764	-0.848	0.488	-0.185	-0.807	-0.473
MeanmeanSAdr	0.327	-1.000	-0.026	-0.661	-0.393	-0.857
MeanmedSAdur	0.211	-0.995	-0.146	-0.747	-0.279	-0.912
MeanpctdurSA	0.347	-0.999	-0.006	-0.646	-0.411	-0.846
MeangrateNA	-0.190	0.993	0.168	0.761	0.258	0.921
MeanpctdurNA	-0.522	0.972	-0.188	0.485	0.581	0.726
MeanglncesMR	0.898	-0.694	0.685	0.059	-0.927	-0.245
MeanduratMR	0.802	-0.814	0.541	-0.124	-0.842	-0.418
MeanmeanMRdr	0.227	-0.997	-0.130	-0.736	-0.295	-0.906
MeanmedMRdur	0.071	-0.971	-0.284	-0.833	-0.141	-0.961
MeanglncesNR	0.998	-0.368	0.911	0.434	-1.000	0.142
MeanduratNR	0.996	-0.387	0.902	0.416	-1.000	0.123
MeanmeanNRdr	-1.000	0.288	-0.943	-0.509	0.996	-0.226
MeanmedNRdur	-0.951	-0.002	-0.999	-0.737	0.927	-0.498
MeansdNRdur	-0.704	0.892	-0.408	0.272	0.752	0.550
MeangrateNR	-1.000	0.335	-0.925	-0.466	0.999	-0.177
MeanpctdurNR	-0.998	0.359	-0.915	-0.443	1.000	-0.152
MaxTdur	0.968	-0.535	0.818	0.259	-0.983	-0.044
MaxRDdur	0.968	-0.535	0.818	0.259	-0.983	-0.044
		Highlights + correlations >.707				Highlights - correls < -.707
		Highlights + correlations >.665, p<.05				Highlights - correls < -..665

To review, the findings from correlational analyses between eyeglance metrics and other driving performance metrics (of lanekeeping, speed keeping, and event detection) can be summarized in terms of several major clusters of effects:

- Metrics associated with glances to the road and SA areas (mirrors and speedometer) tended to be related to lanekeeping (Median SDLP) and to speed keeping (Median Speed Diff). The nature of these relationships was surprising. It was not the case that the more glances to the road, the better the lanekeeping or the less the SDLP. Rather, the more glances to the road, the greater the standard deviation in lane position and the greater the speed difference. It appears that this may reflect some shift of attention from "just lanekeeping and speed keeping" to something else. One plausible explanation was a shift of attention to active monitoring of the roadway, perhaps for event detection, in addition to performance of in-vehicle tasks. Alternatively, longer duration tasks may have been associated with laxer vehicle control because of workload effects (from subsidiary tasks), lost-in-thought distraction effects (for Just Drive), or the continuous effort required for crisp vehicle control. More glances to mirrors were related to increased speed difference.

  Note: Among the road-related glance metrics, there was one exception. Glance Rate to the Road was positively related to missing CHMSLs and negatively related to missing LVD events.

- Metrics associated with glances to task-related areas were related to miss rates for events such as Mean Glance Rate to Task-Related Areas and Percent of Task Duration spent viewing task-related areas. Two were also associated, at a moderate level, with excursions from the lane (% Trials with a Cross of the Lane Line): Mean Number of Glances to Task Related Areas and Mean Total Glance Time to Task-Related Areas.
- For all tasks taken together (Table 3-6), metrics associated with glances to mirrors and situation awareness areas were negatively correlated with CHMSL miss rates and FVTS rates, as were metrics associated with glances to mirrors only. Positive correlations were found for task-related, non-road, and road measures with respect to CHMSL miss rates
- Visual-manual tasks, analyzed as a set (Table 3-7), showed a significant negative correlation between FVTS miss rate and mean single-glance time to the mirrors as well as mean single-glance time to SA locations (perhaps because of the high correlation between SA and Mirrors alone). As the visual dwell time to the mirrors increased on average, FVTS miss rates decreased. The only other significant correlation (positive) involved Mean Glance Rate to the Not Attributable (NA) areas. This is not interpretable because the NA category was an unknown location.
- Auditory-vocal tasks, analyzed as a set (Table 3-8), showed several significant correlations. Negative correlation between FVTS miss rate and mean number of glances and total glance time to mirrors. Significant negative correlations were also found between FVTS and non-road glance counts, total glance time, mean single-glance durations and percent duration. The latter may arise because glances away from the road in these tasks were largely to the mirror events, the latter of which appeared in the left outside mirror. For example, total glance time to mirrors and duration of glances on mirrors, was correlated with missed CHMSL and FVTS events. However, the correlations were negative; as glance durations and time on the mirrors goes down, the miss rate goes up.

Based on these findings, along with an analysis of discriminability, it is possible to make some recommendations about which eyeglance metrics are most useful in future applications:

- For visual-manual tasks:
  - **Traditional measures of eyeglance behavior were confirmed** as being repeatable, having predictive validity, and being useful for discriminating between high and low workload tasks within the visual-manual type of task:
    - Number of Glances to the Task (Mean)
    - Total Glance Time to the Task (Mean)
      - A traditional measure that did meet the criteria for repeatability, but was not highly correlated with other driving performance measures, and did not discriminate between high and low workload tasks within the visual-manual type of task, but is still recommended for use is:

- Duration of Glances to the Task (Mean)
  - It is suggested for retention because it is still included in current practices and, more importantly, because glance duration is controlled to some extent without conscious awareness and, in response to stimulus content, so may be subject to lengthening without awareness in response to visually-demanding stimuli while performing a task and Maximum Task-Related Glance Duration was correlated with lane excursions (Percent of Trials with Cross of the Lane Line). In addition, shorter duration tasks may be associated with typical glances that are longer in duration.

- **Additional metrics provided insight.** They were useful for comparing across tasks of different types and lengths, were repeatable, and prominent in correlations between eye data and performance data. However, they could not make discriminations between tasks of differing workload levels as defined in this study.
- These may nonetheless be useful in research applications, but should be used with caution if applied to product development.

- Proportion of Task Spent Viewing Task (especially as compared to Proportion of Task Spent Viewing Road, which is recommended in its own right)
- Proportion of Task Spent Viewing Mirrors)
- Maximum Glance Duration To Task (Mean)
- Glance Rate To All Areas(Mean)

- For auditory-vocal tasks:
  - No measures of eyeglance are recommended for use with auditory-vocal tasks, since not enough glances to these task types generally occur to make them cleanly interpretable as stand-alone indicators.
  - For research purposes, several metrics provided insight into the possible subtle intrusion of these tasks on driving performance, which was more subtle in its effects than visual-manual demand. However, these glance metrics must always be used with other converging measures of workload (such as event detection) if they are to be interpretable as indicators of workload. If used in this context, the following may be useful in research applications, but should be used with caution:

- Duration of Glances to the Road (Mean)
- Proportion of Task Spent Looking at Road
- Proportion of Task Spent Looking at Mirrors/SA Areas

3.4.4 Summary of Findings from Test Track Eyeglance Data

The key findings of the eyeglance data analyses are summarized as follows:

1. Several categories of eyeglance measures proved reliable in split-half analyses of the data:
   - Number of glances:
     - To road locations
     - To situation awareness locations
     - To task-related areas
     - To Total/All, and To Not Road (combines everything other than Road)
   - Durations of glances:
     - Mean (for most locations: Road, Situation Awareness, Task (was borderline), and Not Road
     - Median (for some locations: Road, Situation Awareness)
     - Standard deviation (for some locations: Road, Task, Total/All, Not Road)
     - Max (for only certain location types: Road, Task, Total/All)
   - Accumulations of durations for certain location types:
     - Total Glance Time to Road Location
     - Total Glance Time to Situation Awareness Location
     - Total Glance Time to Task-Related Areas
   - Percent (or proportions) of task time spent looking at a location type:
     - To Road Locations
     - To Situation Awareness Areas (borderline)
     - To Task-Related Areas
   - Rates of glances per second
     - Overall (Total/All), Road, Task, Not Road
2. These same measures tended to reveal interesting findings. First, and very important among these findings, was the fact that not all information is in the simple classification of glances as on-road or off-road. Glances to road, task, and mirror locations all carried important information. Among these, there were measures that discriminated between types of tasks. There were distinct patterns of glancing revealed across types of locations (road, mirrors, task).
3. Among the most interesting findings from formal statistical analysis was a significant Task by Location Type interaction across many of the eyeglance measures. Notable was the fact that the pattern of glances to the roadway discriminated task types particularly well, and a measure that integrated multiple measures together—proportion of task time spent looking at the road (Pct Dur Rd)—was particularly useful for characterizing patterns of glancing associated with tasks, along with a similar measure applied to each other glance location (task and mirrors).

In summary, there were multiple effects of in-vehicle tasks on eyeglance behavior. Eyeglance metrics showed distinct patterns for different types of task engagement (just driving versus concurrently performing an auditory-vocal task or concurrently performing a visual-manual task). The Just Drive task was distinguished by patterns in which drivers looked at the road about 83 percent of the time and scanned their mirrors about 14.3 percent of the time. Glances on the road were about 8 seconds duration, on average. Auditory-vocal tasks showed a somewhat similar pattern, though drivers gazed at the forward roadway somewhat more (88%), using longer gazes (9 to 16 seconds, on average), and scanned their mirrors somewhat less (11%). The miss rate for event detection was slightly elevated over just driving for auditory-vocal tasks for CHMSL and LVD events, showing an increase of ~4 percent for CHMSL and LVD events, and somewhat more for peripheral FVTS events, showing an increase of ~23 percent, although event detection was less affected by auditory vocal tasks than by visual-manual tasks. Visual-manual tasks showed a different pattern, in which drivers looked at the forward roadway much less (viewing the road only 34 to 61 percent of the time during a task, and using glance durations on the road that were less than 2 seconds, on average. This reduction in glances to the road was made in order to view task-related areas required for performing the in-vehicle activity (viewing the task 29 to 60 percent of the time during its length). For visual-manual tasks, glances tended to cycle frequently back-and-forth between the task and the roadway locations, and glance rate measures proved to carry interesting information. Visual-manual tasks led to a more pronounced reduction in mirror-scanning (to 7%) and were associated with higher rates of missed events, although this was sometimes due to a methodological constraint for LVDs. Increases in miss rates over Just Drive were approximately 14 percent for CHMSLs, 20 percent for LVDs, and 42 percent for FVTS events, on average.

4. Interrelationships with driving performance measures revealed:
    - Correlations with SDLP
    - Correlations with Speed Difference
    - Correlations with Event Detection (due to influence of Just Drive and selected tasks)
5. A striking new finding emerged from relating eyeglance data to event-detection data. Qualitative exploration of the time series data suggested that event detection affected eyeglance behavior. In brief, formal analyses of task summary statistics (a detailed analysis of the time series will be a target of future work) indicated that when an event occurred and was responded to, eyeglance behavior changed such that:
    - For CHMSL events,
        - Durations of glances decreased slightly for all locations except to situation awareness locations
        - Rate of glancing increased slightly to road and situation awareness areas (mirrors)
    - For LVD events,
        - Durations of glances to the road lengthened
        - Rate of glancing decreased to task-related and situation awareness areas
    - For FVTS events,
        - Durations of glances decreased

- Rate of glancing to road and situation awareness areas increased

Changes to glance durations interacted with Task Type and were more pronounced for Just Drive and auditory-vocal tasks than for visual-manual tasks, which usually showed a different pattern. Events, when detected, appeared to act as attentional interrupts for auditory-vocal tasks and the Just Drive tasks, in eliciting more active scanning of the forward roadway and mirrors. This was a strategy that would be expected to improve subsequent event detection. Event detection also affected glance behavior during visual–manual tasks, but somewhat differently. Rate of scanning between all locations (road, task, and mirrors) increased, but higher glance rates were associated with higher rates of missed events (except for LVD events).

The finding that event detection may affect glance behavior has implications for analysis and design of future studies. Methods used to study event detection may influence the behavior of interest and suggest that when evaluating the visual demand of tasks in an advanced information system or in-vehicle device, it is important that multiple test trials be conducted—some with and some without event detection. The trials used to evaluate the visual demand of a task should not include events to-be-detected in order to obtain clean measurements of glance behavior, free from the influence of co-occurring events.

The findings on event detection, substantive and methodological, highlight a rich area for future exploration.

6. Recommendations on eyeglance metrics for use in future work. The usefulness of the traditional eyeglance metrics for visual-manual tasks was confirmed through analyses of repeatability, predictive validity, and discriminability. These included: number of glances to task-related areas and total glance time to task. Retention of glance duration of task-related glances was recommended as well. Additional metrics, which emerged from new findings from this research, were recommended for use in future research on visual-manual tasks. No eyeglance metrics were recommended for application to the assessment of auditory-vocal tasks, although for research purposes, metrics emerging from this work as promising were identified (i.e., Proportion of task duration spent looking at the road, and mean duration of glances to the road, and proportion of task duration spent looking at mirrors/situation awareness areas).
7. Eyeglance behavior appears to be a key diagnostic for workload, and its associated metrics offer promise as key discriminators in identifying tasks that interfere with visual performance on the road.

3.5 Test Track Task Effects on Lateral Control

Lateral control is safety-relevant. Lane departures are the first critical event in single-vehicle road departure crashes, lane change crashes, and opposite direction crashes. Jointly, these crash types represent a substantial portion of the crash problem. The intersection of these types of crashes with driver distraction causal factors provides the motivation for measuring and analyzing lateral control performance.

Many different measures of lanekeeping can be defined. In the DWM project, two measures have been selected for in-depth evaluation. The rationale behind these selections is provided below.

Standard Deviation of Lane Position (SDLP) is defined as the square root of the average squared deviation in lane position about the mean lane position observed during a task. It is measured in feet, and was obtained with an Assistware Technology, Inc. lane tracker. SDLP was

calculated only if valid lane tracker data were available for 85 percent or more of a task trial duration. For each participant, each task trial generated an SDLP value. For each participant, the SDLPs were averaged over his or her replications of a given task. This created a single SDLP value for that participant performing that task. If only a single trial was available for a participant, then its SDLP value was used for that task from that participant. The median of all such SDLPs for a given task was used as a summary statistic for that task. The median was chosen because it is more resistant than the mean to outliers or extreme values in the data. Its use avoids the need for data truncation or data transforms to approach normality. The median can provide a robust "typical" value of task effects.

Percent Lane Exceedance(Cross) Trials is the percentage of participants who had one or more lane exceedances during one or more trials for a given task. A lane exceedance (cross) event was defined to have occurred if the leading edge of the subject vehicle crossed the adjacent lane line's outer edge. For each participant, each task trial generated a Lane Exceedance count. These counts were converted to a binary score, one (1) if there were one or more lane exceeds, zero (0) otherwise. For each participant, these binary scores were averaged over his or her replications for a given task. This created a single Lanex (Cross) Trials value for that participant performing that task. If only a single trial was available for a participant, then its Lanex (Cross) Trials value was used for that task from that participant. The average of these values was calculated and multiplied by 100 to create Lanex (Cross) Trials percentages. Lane exceedances were manually verified by staff through review of track pavement video to confirm lane tracker output. This manual method did not allow for measuring the lateral extent of a lane exceedance. The lane tracker did not always provide reliable data on the lateral extent of a lane exceedance if the lane tracker lost the lane line(s). Lateral extent of a lane exceedance was not assessed for this reason. Finally, the percentage of participant cross trials was used in lieu of lane exceedance counts because test participants did not all complete the same number of trials per task.

These two measures provide complementary data to distinguish task effects. Lane exceedances are discrete and infrequent events that can provide an indication of degraded lanekeeping. Lane exceedances are also of procedural interest. From a practical standpoint, lane exceedance event counts and durations out-of-lane can be captured through video and can be manually reduced and verified. Manual review of video, however, does not allow accurate assessment of the lateral extent or overridden area.

SDLP on the other hand, is a continuous, ever-present measure of lanekeeping. Normal probability theory suggests that larger SDLP implies an increased likelihood of departing the travel lane eventually (Allen, Parseghian, and Stein, 1996). While there may be no lane exceedances for a given trial, there is always lanekeeping to measure. However, SDLP requires a lane tracking system for data capture. Such systems, like eye trackers, are at least sometimes less than robust. They are also expensive.

Lateral control usually is more sensitive to task effects than longitudinal control because lateral position can change more quickly. It has proven useful in studies of driver distraction, drowsy driving, and intoxication. Police use lateral control as an indicator of impaired driving because, over years of real-world experience, it is relevant. On the other hand, measures based on steering inputs have not been used in this analysis because of the noisiness of steering data in the face of road and vehicle characteristics, and the more remote association to lanekeeping. However, measures based on steering inputs have been used successfully for over 30 years (e.g., McLean and Hoffman, 1975) and merit future evaluation.

3.5.1 Standard Deviation of Lane Position (SDLP)

Figure 3-40 presents the median SDLP for the 23 tasks evaluated on the test track. The range of median SDLP values was between 0.45 ft and 0.8 ft, within the normal range. The data are

ordered in such a fashion that task duration may play a role in interpretation of the results. For example, Just Drive had greater median SDLP values than almost all other tasks except the two most visually demanding tasks performed on the track (Destination Entry and Route Tracing) and the Delta Flight Information mixed-mode task.

For reference, Figure 3-41 provides the median track task durations for all tasks. Auditory-vocal tasks (except Book-on-Tape Summarize) and Just Drive were of fixed durations (approximately 2 min). Visual-manual tasks were duration-intrinsic. That is, visual-manual tasks were not of fixed duration, they took as long as the test participant spent to complete them.

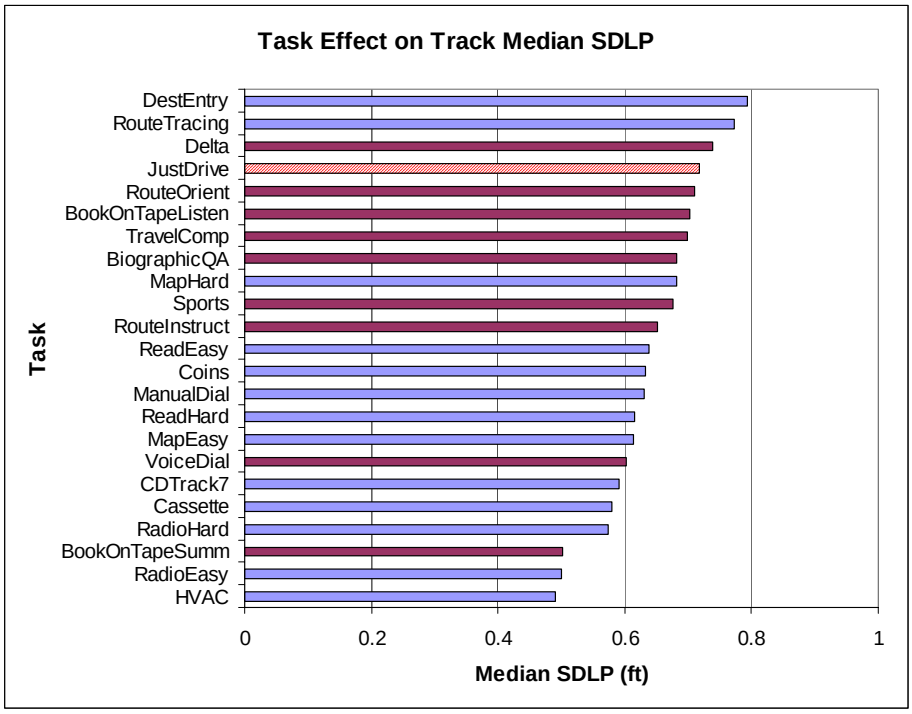


Figure 3-40. Track Median Standard Deviation of Lane Position (SDLP) by Task

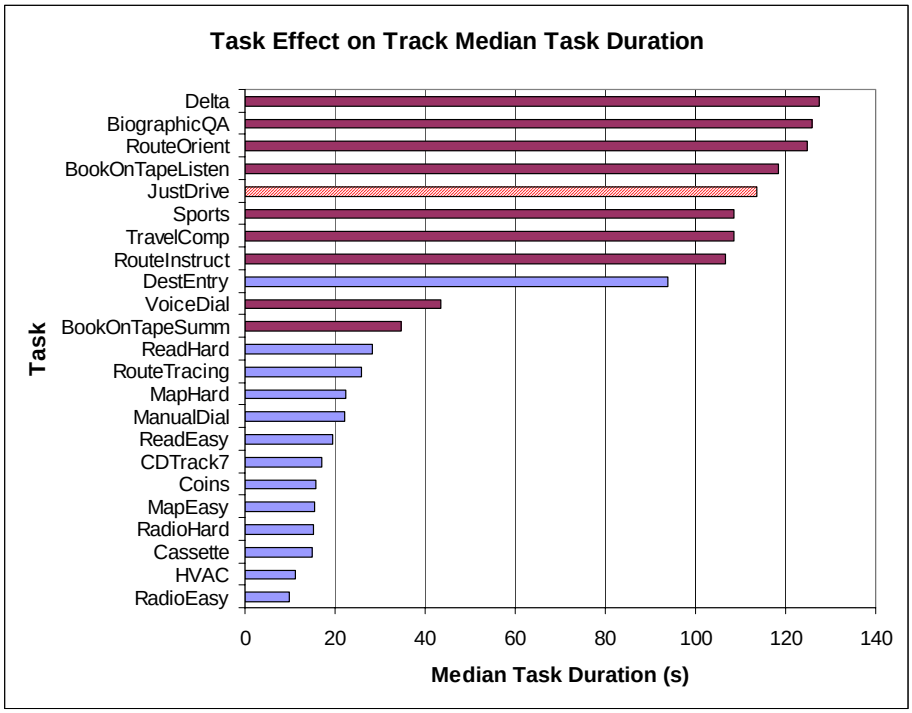


Figure 3-41. Track Task Effects on Median Task Duration

Figure 3-42 is a plot of Median SDLP versus Median Task Duration for visual-manual tasks only. The visual-manual tasks clearly showed the effect of task time on the results. Destination Entry was a standout (far right point on graph), extreme in both typical duration and in typical SDLP values. Interestingly, Route Tracing had a median SDLP value similar to Destination Entry, yet its typical task time was almost four times smaller (about 26 seconds versus 94 seconds). It appears that the participants would typically not allow continuous lanekeeping variability to increase without limit, even with a very demanding task like Destination Entry or Route Tracing. (The effects of these tasks in terms of lane exceedances may shed light on this phenomenon). Overall, the main trend was for SDLP to increase with increases in task time.

Figure 3-43 is a plot of Median SDLP versus Median Task Duration for auditory-vocal tasks and Just Drive. Just Drive was included because it was of similar duration to the auditory-vocal study tasks (except for Book-on-Tape Summarize) and, like them, imposed no additional visual input demand or manual output demands. Note that even though the relative range in median task duration for these auditory-vocal tasks was approximately the same as for the visual-manual tasks (a spread of approximately 20 seconds), there was no systematic relationship with SDLP. The standout here was the Book-on-Tape Summarize task, which had a much shorter median task duration than the other tasks and a much smaller median SDLP value. The similarity of median SDLP values for auditory-vocal and Just Drive tasks may be explained by predictions from the literature. Brown (1994) explained that general withdrawal of attention (such as that associated with eyelid closure or eyeglances away from the road scene) can affect vehicle control and object-and-event detection. On the other hand, selective withdrawal of attention can leave over-learned vehicle control performance unaffected. Selective withdrawal of attention is more central or cognitive in nature and is not necessarily associated with significant eyeglances away from the road scene.

These figures indicate that SDLP is associated with the time required to perform visual-manual tasks while driving. The findings suggest that the shorter in duration a visual-manual task is, the smaller the variation in lanekeeping will be during that in-vehicle activity. That is, lanekeeping associated with SDLP is constrained by task duration, at least partly. On the other hand, typical SDLP values are greater for visual-manual and Just Drive tasks. However, there is not much variation in SDLP as a function of duration among the long, auditory-vocal and Just Drive tasks. This latter result holds even though the relative range in durations among the auditory-vocal tasks is about the same as that among the shorter visual-manual tasks, about 30 seconds. But a long visual-manual task like Destination Entry had greater median SDLP than any of the auditory-vocal tasks despite being somewhat shorter. Median SDLP for Destination Entry did not lie along the trend of the other visual-manual tasks. SDLP grows as task time grows, but only up to a certain point.

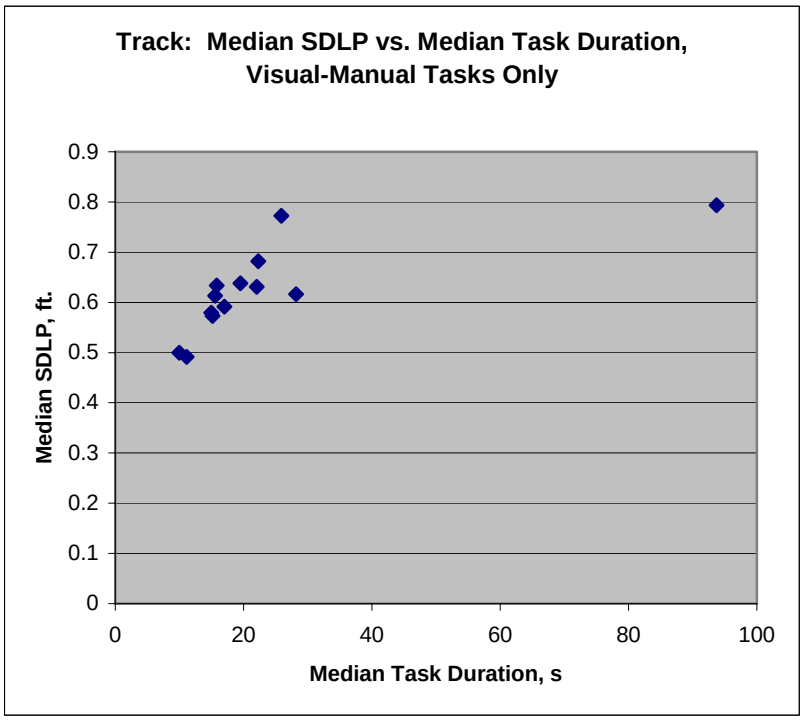


Figure 3-42. Track Median SDLP as a Function of Median Task Time for Visual-Manual Tasks Only

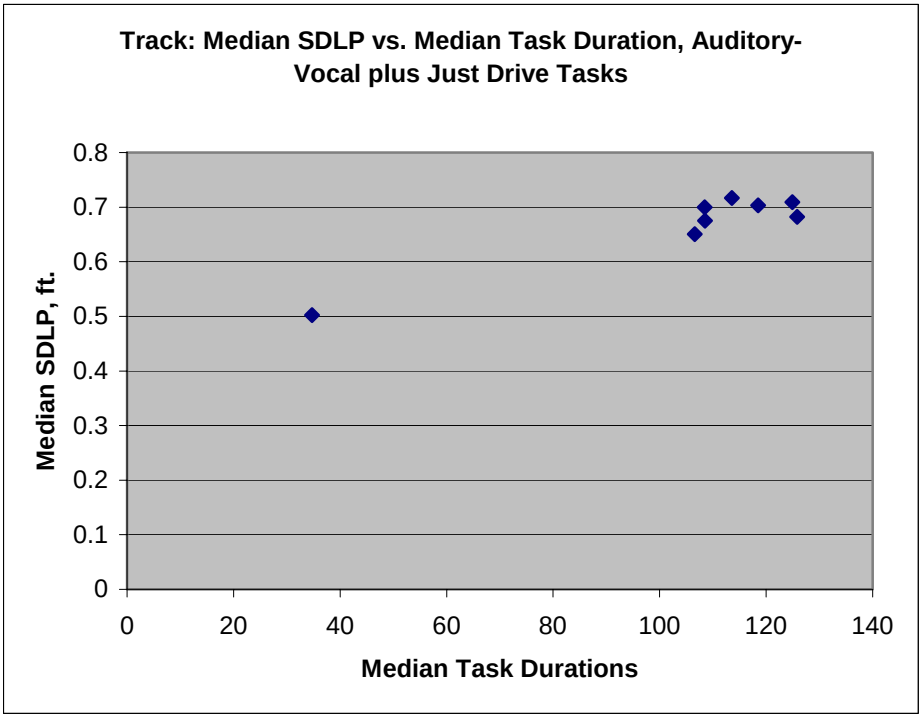


Figure 3-43. Track Median SDLP as a Function of Median Task Time for Auditory-Vocal Tasks

3.5.2 Percent Lane Exceedance (Cross) Trials

Figure 3-44 shows the ranking of tasks in terms of Percent Lane Exceedance (Lanex (Cross)) cases. Recall that this is a measure of the percentage of participants who had one or more lane exceedances during a given task trial. This figure shows Destination Entry and Route Tracing were head and shoulders above the other tasks. The ordering of tasks does not seem as clearly related to time as SDLP.

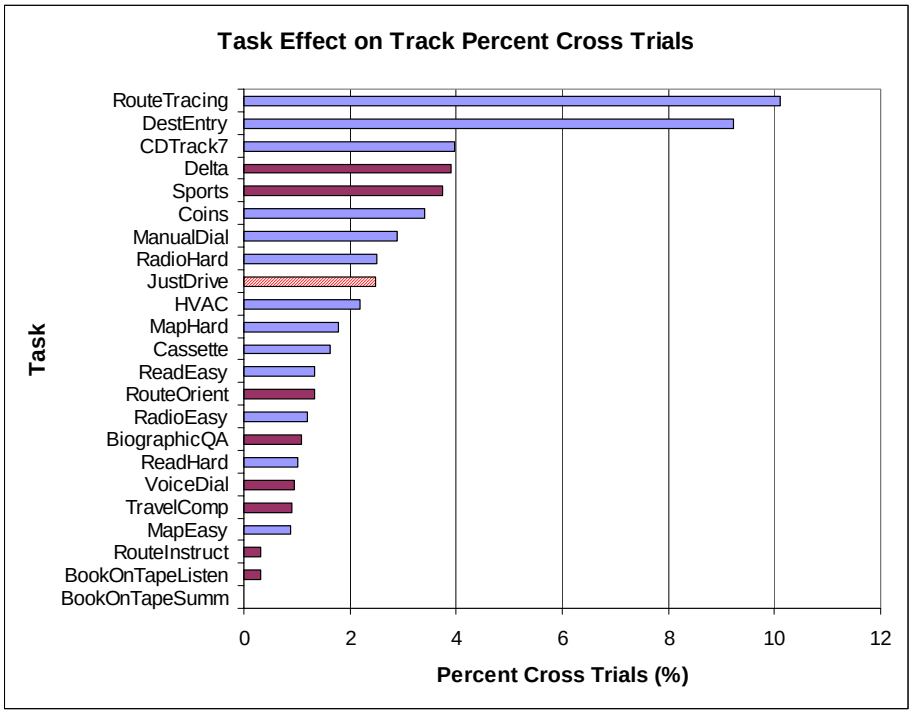


Figure 3-44. Track Percent Lanex (Cross) Cases: Track Results

Figure 3-45 is a plot of Percent Lanex (Cross) as a function of Median Task Duration for visual-manual tasks. There does not appear to be any relationship between Task Duration and the occurrence of lane exceedances in this data. The two extreme points were Route Tracing and Destination Entry. Lane exceedances did serve a useful purpose in indicating high levels of visual-manual demand. The Lane exceedance results for these to tasks were consistent with the ranking of these tasks based on typical SDLP.

Figure 3-46 is a plot of Percent Lanex (Cross) as a function of Median Task Duration for auditory-vocal tasks plus Just Drive. There is no relationship between the two. Also, the Book-on-Tape Summarize had no lane exceedances, despite a median duration comparable to visual-manual tasks that did exhibit lane exceedances.

Taken together, lane exceedances appear to be unrelated to task duration. They arise in visual-manual tasks with high demand. They are not systematically related to other tasks, at least based on prior predictions. They are also infrequent. Lane exceedance data should be sought and is appear suitable to detect very high workload effects.

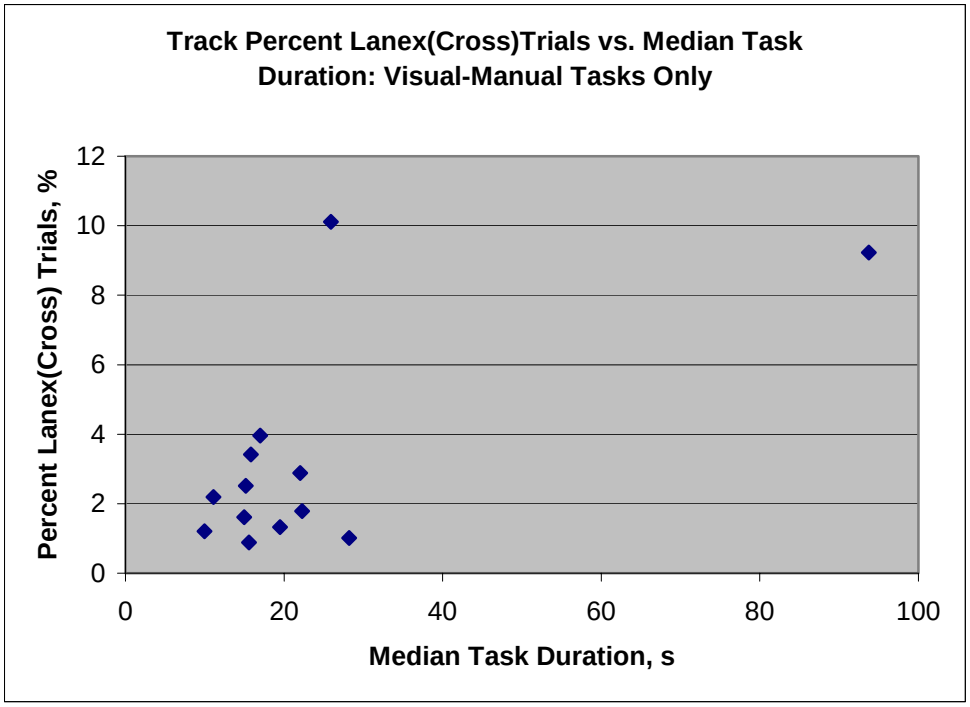


Figure 3-45. Track Percent Lanex (Cross) Trials by Task Duration: Visual-Manual Tasks Only

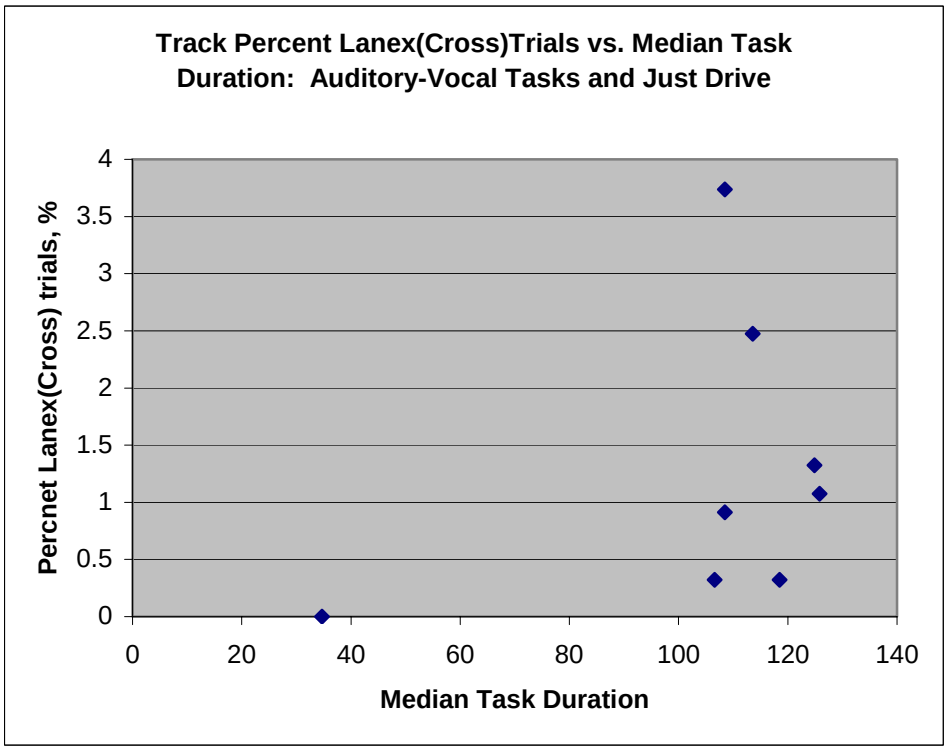


Figure 3-46. Track Percent Lanex (Cross) Trials Versus Median Task Duration: Auditory-Vocal Tasks and Just Drive

In closing this section, a comment on the repeatability of the median SDLP and Percent Lanex (Cross) trials is appropriate. Repeatability was assessed using a split-group approach. Split-group repeatability analysis addresses the question: "Would I get similar results if I ran the study on a different group of people?" To answer this question, participants in each venue were randomly assigned to one of two groups. The two groups were roughly equated for age and gender. Next the summary statistics for each of the groups were calculated for each of the tasks. Then the task summary statistics (23 tasks) were regressed on one another. A correlation of r = 0.70 or greater was taken as indicative of repeatability. The choice of this criterion was based on the logic that approximately 50 percent (r-squared) in the variability of one group's values could be accounted for by the variability of the other group's values.

Table 3-10 contains the Split-Group Repeatability results for selected track measures, including SDLP and Percent Lanex (Cross) Trials. Track median SDLP results indicated a split-group correlation of approximately 0.82 between the two groups' outcomes. A similar analysis of Track Percent Lanex (Cross) trials yielded a correlation between groups of only about 0.63. Exploration into lane exceedance duration indicated that this measure was highly unreliable and so was dropped early on. In all, the track repeatability results for this study indicated good repeatability for SDLP measurements but marginal repeatability for the lane exceedance measures. Percent Lane Exceedance (Cross) Trials, however, was robust in identifying high demand associated with Destination Entry and Route Tracing.

Table 3-10. Repeatability of Selected Test Track Driving Performance Measures

Driving Measure	Split Group Level Correlation, r	Split Group R^2 %	Estimated Stdev about regression line, S	P-Value, Sig Value
Mean Task Duration	0.999	99.8	1.858	0.000
Median Task Duration	0.999	99.8	2.304	0.000
Mean SDLP	0.912	83.2	0.034	0.000
Median SDLP	0.822	67.6	0.045	0.000
Mean Speed Diff	0.938	88.0	0.445	0.000
Median Speed Diff	0.942	88.7	0.456	0.000
Pct Cross Trials	0.631	39.9	1.803	0.001
Mean Cross Duration	0.322	1.3	3.365	0.134
Median Cross Duration	0.322	0.9	3.399	0.134
Pct LVD Miss Rate	0.733	53.7	13.180	0.000
Mean LVD RT	0.536	28.8	0.569	0.008
Median LVD RT	0.553	30.6	0.765	0.006
Pct CHMSL Miss Rate	0.753	56.7	6.293	0.000
Mean CHMSL RT	0.393	15.4	0.189	0.064
Median CHMSL RT	0.639	40.8	0.169	0.001
Pct FVTS Miss Rate	0.782	61.2	6.537	0.000
Mean FVTS RT	0.432	18.6	0.244	0.040
Median FVTS RT	0.420	17.7	0.322	0.046

3.6 Test Track Task Effects on Longitudinal Control

Longitudinal control is critical to maintaining vehicle separation. With degradation in longitudinal control, such as decreased range and increasing range rates, the potential for rear-end collision increases. Rear-end collisions account for a large number of accidents and systems to prevent dangerous ranges or closing rates are studied extensively. Thus forward range and range rate are important metrics to examine for potential effects of driver distraction.

Another longitudinal metric is vehicle speed. Accidents caused by large variances in speed occur both in low visibility and dense traffic situations. Speed is also often a factor in roadway departure accidents that occur on curved sections of roadway.

Numerous measures of longitudinal control were examined in the DWM study. These measures include forward range, range rate, speed, and time headway. Measures of variance and central tendency such as minimum, mean, median, maximum, and standard deviations can be calculated. For this study, measures of range, range rate, and speed were selected for in-depth analysis.

In the DWM study, the vehicles driven by test participants were equipped with Delphi ACC1 forward range sensors. The sensors were modified to output information on the range in feet, range rate in feet per second, and lateral location of a vehicle ahead of the subject vehicle. Data quality data standards required at least 90 percent of each individual task performance in order for any range data to be included in analysis. Speed of the subject vehicle is calculated from the OEM transmission sensor and recorded in feet per second.

For analysis of longitudinal measures, task performances were averaged across all replications of a task that did not contain a lead vehicle deceleration event for each subject. All tasks of a particular type, visual-manual, auditory-vocal, Just Drive, and mixed-mode were then averaged across tasks and subjects. For this analysis, the mixed-mode tasks, containing both visual-manual and auditory-vocal components were grouped separately into the combo task type. Just Drive was grouped separately from the other task types. This data was used as the input to analysis of variance to examine potentially significant task effects on vehicle control. While the results of this analysis are mentioned where appropriate, all graphs in this section will present data by individual task. The task data were averaged across all tasks that did not contain a lead vehicle deceleration event for each subject, as mentioned earlier. Data were then averaged across subjects to yield a mean performance metric inclusive of all subjects for each task.

3.6.1 Minimum, Mean, and Maximum Measures

Figure 3-47 shows the mean values of the minimum, mean, and maximum range to the lead vehicle for each of the 23 test track tasks plotted against task duration. Table 3-11 presents the list of numeric codes used in Figure 3-47 and several other figures in this section. Shorter tasks (almost all visual-manual tasks and the Book-on-Tape Summary task), both the conventional visual-manual tasks and those requiring paper stimulus materials, are closely grouped for all three measures. The other large cluster contains the longer duration auditory-vocal tasks and Just Drive task. On the edge of the visual-manual cluster is the Voice Dial (mixed-mode) task, also a shorter-duration task. Destination Entry, with a much-longer task duration than the other visual-manual tasks, lies in the graph near the auditory-vocal cluster. This is also true for the longer duration mixed-mode Delta Flight Information task. While the mean range varies only by approximately 14 feet across all tasks, maximum and minimum ranges vary by 20 and 22 feet respectively. This spread seems to be mostly an effect of task duration as the longer auditory vocal tasks are at the more varied end of the ranges. Destination Entry, as it often does, stands out as significantly different from the other visual-manual tasks.

An analysis of variance shows that there are statistically significant differences in each of these measures among different task types. Minimum range shows the most differentiation between tasks, often along the visual-manual versus auditory-vocal division.

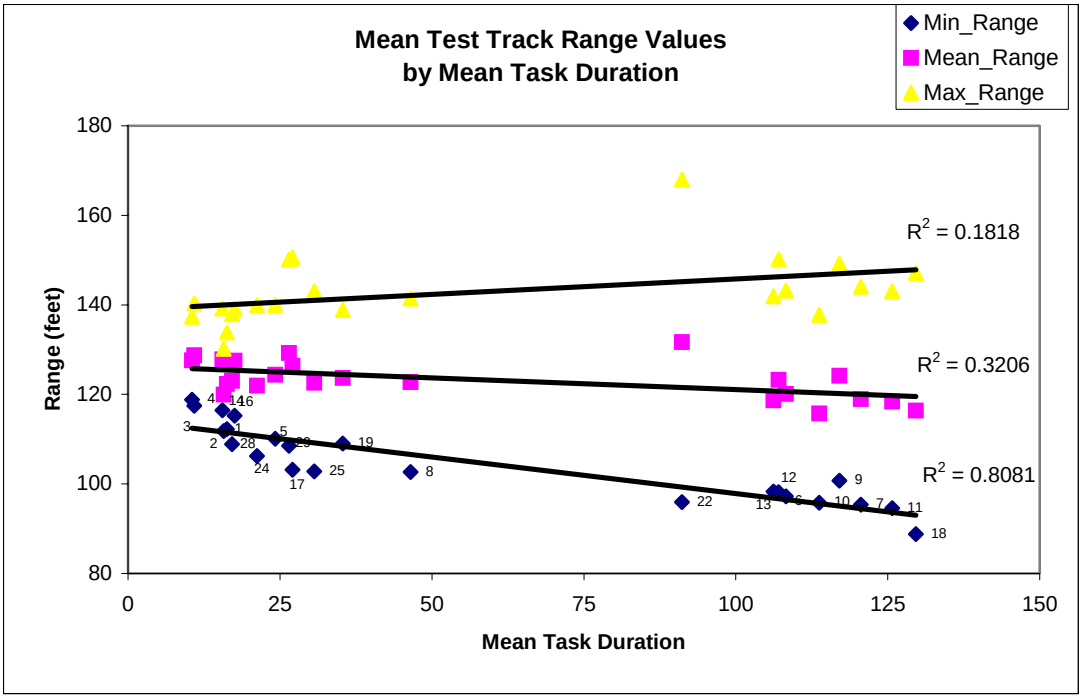


Figure 3-47. Mean Test Track Longitudinal Metrics by Task Duration

Table 3-11. Numeric Codes Assigned to Tasks

Numeric Code	Task Name	Numeric Code	Task Name
1	Coins	13	Sports Broadcast
2	Cassette	14	Radio Tune Hard
3	HVAC	16	CD/Track 7
4	Radio Tune Easy	17	Route Tracing
5	Manual Dial	18	Delta Flightline
6	Travel Computations	19	Book-on-Tape Summary
7	Route Orientation	22	Destination Entry
8	Voice Dial	24	Read Text Easy
9	Book-on-Tape Listen	25	Read Text Hard
10	Just Drive	28	Read Map Easy
11	Biographical Q&A	29	Read Map Hard
12	Route Instructions		

Figure 3-48 shows the mean range values for test track tasks. This graph shows the difference in the spread of range values for shorter visual-manual tasks as compared to the longer visual-manual and auditory-vocal tasks. Destination Entry, the longest visual-manual task, shows the most variation in range. Destination Entry was shorter in duration than the longer duration auditory-vocal tasks, Delta Flight Information, and the Just Drive tasks, which follow it in range variability. It is also interesting to note that Just Drive, a fixed duration task, has more range variability than most of the visual-manual tasks. While statistically significant differences between task types exist, the main difference is that with the shorter visual-manual tasks time for variation in longitudinal position is limited. The positions of the visual-manual Destination Entry task, the mixed-mode Delta Flight Information task, and the auditory-vocal Book-on-Tape Summarize task indicate that longitudinal variability is at least partly a task duration dependent phenomena.

Figure 3-49 shows the mean range rate values for all tasks. Mean range rate shows little variability across tasks, all near zero. The more interesting measures are minimum and maximum range rate, especially the latter. With these measures, in general, a slight trend toward less stable car following can be seen increasing from the shortest visual-manual tasks and auditory-vocal Book-on-Tape Summarize task through the longer auditory-vocal tasks. Maximum range rate grows more starting with the Delta Flight Information task through the most demanding visual-manual tasks. Surprisingly, Book-on-Tape Listen is also in this group of higher closing rate tasks. Shorter visual-manual tasks are associated with smaller extremes. However, the greatest extremes are associated with other visual-manual tasks, many of which are predicted to be higher-workload tasks based on prior prediction.

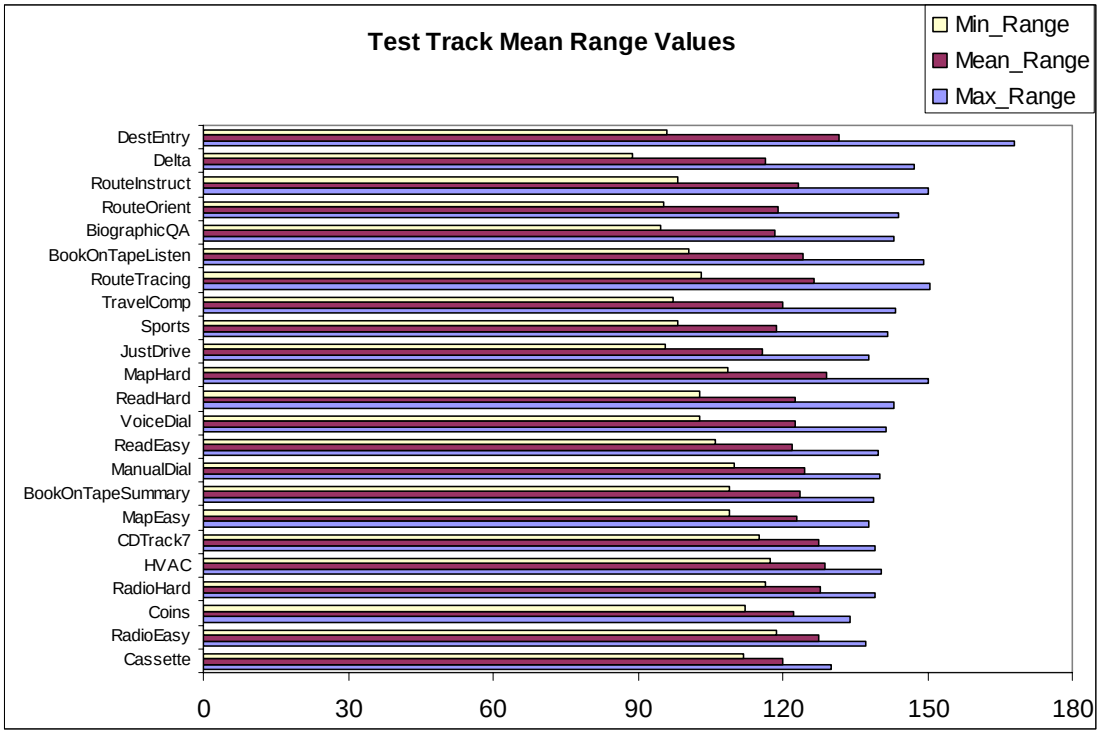


Figure 3-48. Mean Test Track Range Values

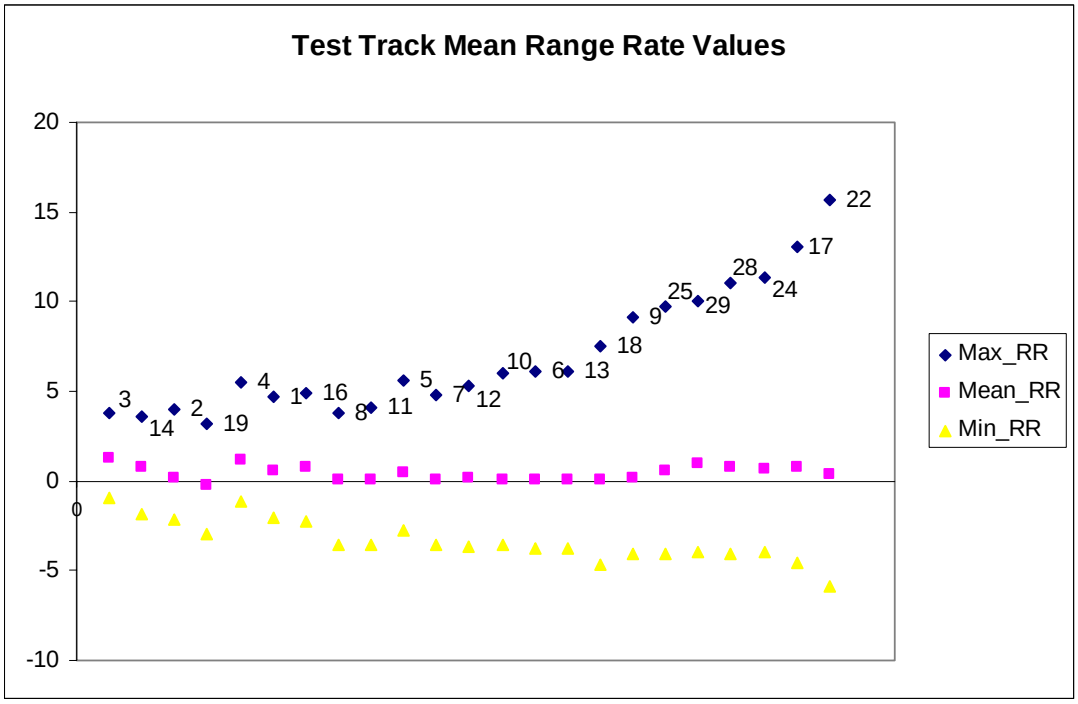


Figure 3-49. Mean Test Track Range Rate Values

Figure 3-50 presents the mean values of speed for all tasks. This graph shows, that in general, the longer tasks allow more time for variation in vehicle speed. Destination Entry, Delta Flight Information, Route Tracing, and Read (Hard), all shorter in duration than the fixed length auditory-vocal tasks, show the most variability in speed. As with range rate, this mixing of short and long tasks with regard to speed variability indicates that this measure of longitudinal control is not solely dependent on time. The differences between tasks, however, are relatively small and practical significance must still be determined.

Figure 3-51 shows Speed Difference, calculated as maximum minus minimum speed by task. This is a clearer picture of speed variability showing the same information discussed above. Analysis of variance showed significant task effects for this metric, though tasks are more interspersed than with other measures.

Figure 3-52 shows the relation between Speed Difference and Task Duration. While longer tasks allow more time for longitudinal position to vary, the correlations in this graph might suggest that it is not the only factor. Much of the correlation shown is due to the two clusters of tasks. Task Duration is highly predictive of Speed Difference for the entire task set. Rather there is a trend seen in short visual-manual tasks that is distinct form that for longer auditory-vocal tasks, indicating a difference that is not attributable solely to task duration.

Figure 3-53 shows Speed Change, calculated as final speed minus initial speed. While the relative differences in Speed Change are small, the analysis of variance showed a significant difference between task types for this measure as well. It can be seen that now, the only visual-manual task being grouped with the auditory-vocal tasks is the longer duration Destination Entry task, as is commonly seen with other measures. On the other hand, the shorter duration Book-on-Tape Summarize task is grouped with the shorter visual-manual tasks. The Just Drive task is also grouped with the shorter-duration tasks even though it is a 2 minute task. This grouping indicates

that, generally, drivers are traveling slightly faster at the end of a longer task (regardless of type) than at the beginning. This pattern is reversed for shorter visual-manual and Book-on-Tape Summarize tasks with drivers traveling slower at the end of the task. This trend would be seen if drivers are shedding car following during tasks, then surging forward to catch up to the lead car once attention is returned to car following either at the end of a short task or during longer tasks. However, this explanation does not account for the results of the 2-minute Just Drive task, for which there was no subsidiary task to shed.

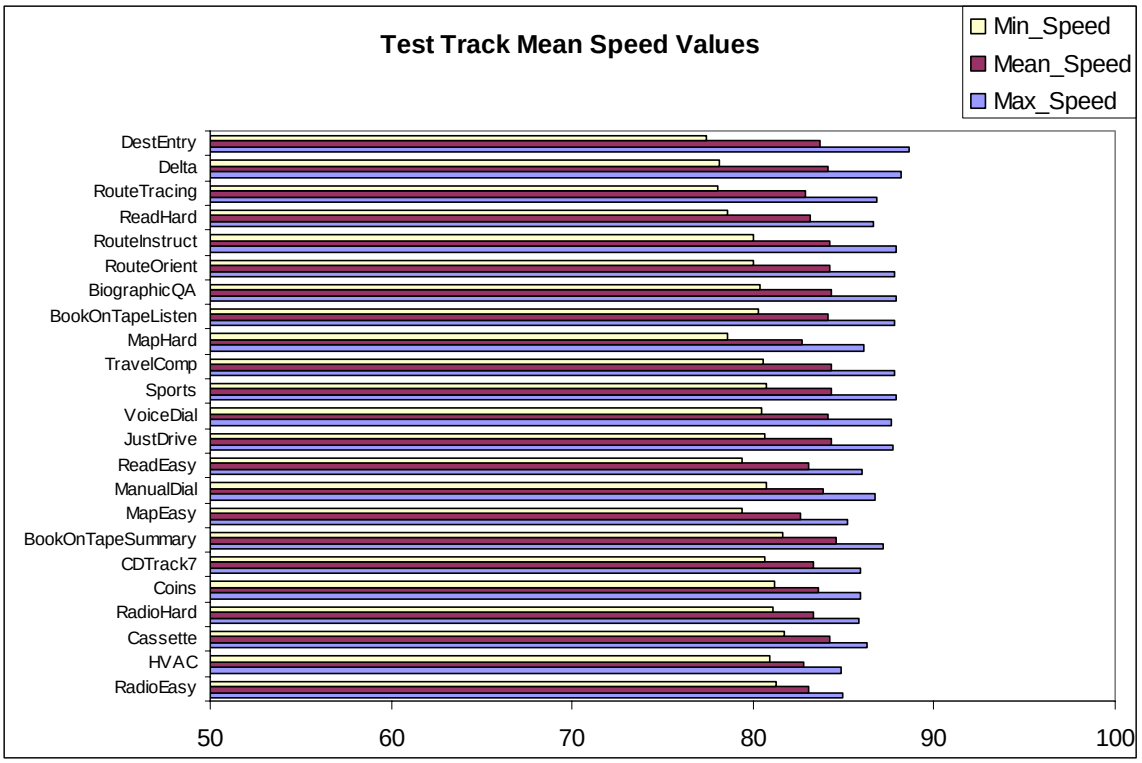


Figure 3-50. Test Track Mean Speed Values

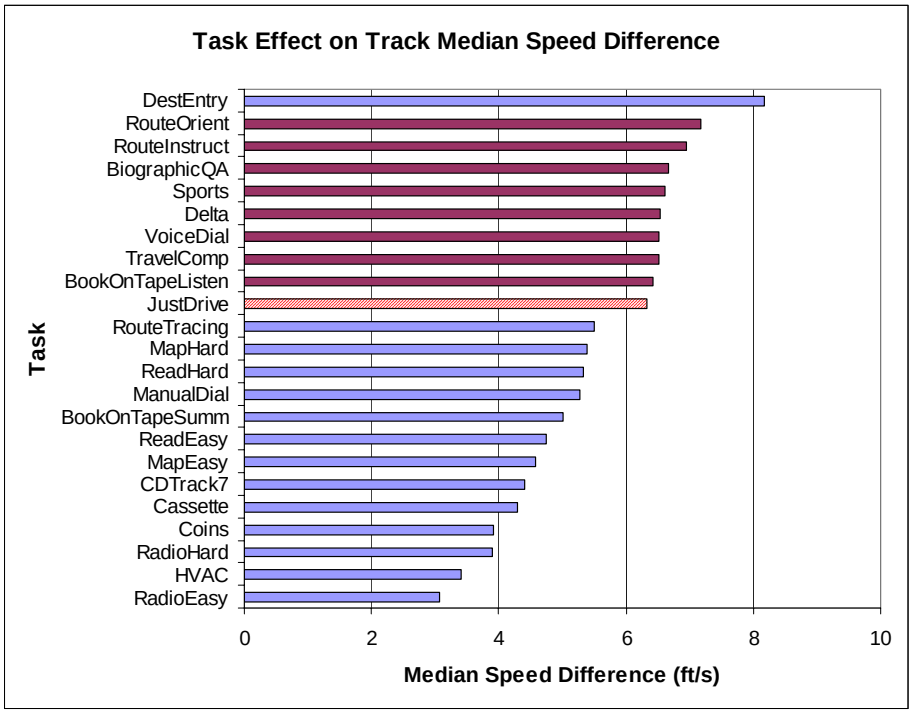


Figure 3-51. Test Track Median Speed Difference Values by Task

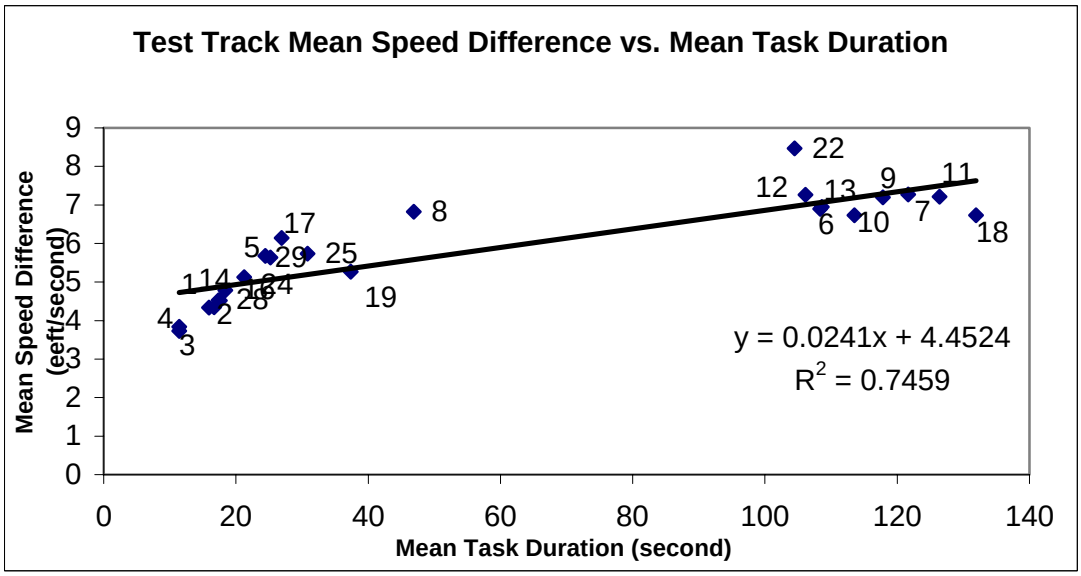


Figure 3-52. Test Track Speed Difference Versus Task Duration

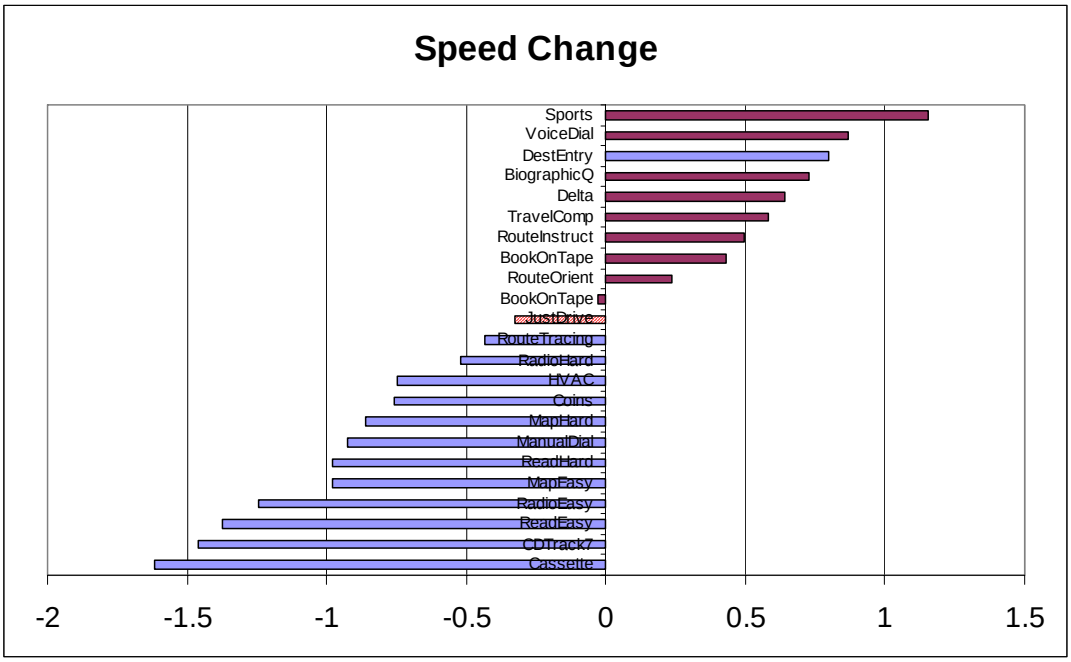


Figure 3-53. Test Track Mean Speed Change Values by Task

3.6.2 Split Group Reliability of Measures

To assess repeatability of longitudinal measures, a split-group repeatability analysis was performed. All test track subjects were divided into two groups of roughly equal distribution of age and gender. All measures were then summarized by task number and by task type within each group. High correlation between the same measures from the two groups was then taken as evidence of repeatability of the metric.

The most repeatable measures (i.e., highest correlations) between groups were for time-related measures such as Time at Minimum Speed at nearly a 0.99 correlation. The correlation between these measures and task duration was just as high, so these measures were not chosen as indicators of workload. Correlations for measures of range, range rate and speed were lower, typically 0.70 to 0.85 with a number of measures having lower correlations between split groups.

When tasks are grouped into the four task types, repeatability correlations rise significantly, as can be seen in Table 3-12. Other measures of longitudinal control were eliminated from consideration due to lower repeatability and/or very high correlation with task duration.

Table 3-12. Repeatability Correlations for Test Track Longitudinal Measures

Test Track Split Group Reliability		
Metric	**Task #**	**Task Type**
Min_Range	0.8044	0.8320
Mean_Range	0.4445	0.7185
Max_Range	0.7060	0.9710
Min_RR	0.7953	0.9121
Mean_RR	0.8465	0.9923
Max_RR	0.4179	0.7885
Min_Speed	0.6024	0.7957
Mean_Speed	0.8497	0.9681
Max_Speed	0.9462	0.9852
Speed_Diference	0.7872	0.9352
Speed_Change	0.9068	0.9464

3.6.3 Summary of Findings from Test Track Longitudinal Metrics

When examining these measures by task or by task type, multiple statistically significant differences can be found. Due to the relative stability of automobiles longitudinally, changes in these measures are somewhat dependent on time. While mean values of range and range rate will be dependent on initial conditions and individual drivers' personal preferences, minimum and maximum values could be a better indicator of the quality of longitudinal control. Similarly, mean speed shows less variation between tasks and it is more informative to examine Speed Difference and Speed Change. Range and range rate difference and change may also be informative but were not examined in this study.

While longitudinal variation is somewhat dependent on time, due to vehicle dynamics, time does not explain all the task effects seen in these metrics. For instance, Speed Difference is correlated to task duration with an R^2 value of 0.75 for the entire task set. This is due mainly to two distinct groupings of tasks, short and long duration tasks. The correlations within each of these two groups indicate different relationships between time and speed between the task types. These correlations, together with the results shown here, indicate that while important, task duration is not the only influence on longitudinal control metrics.

The summary data show that there are tasks with less variability in longitudinal position than Just Drive. These tend to be the short duration visual-manual tasks, which exhibit higher-minimum and lower-maximum measures of range, range rate, and speed. Longer tasks, both auditory-vocal and visual-manual, tend to show more variability than Just Drive and in these comparisons, task durations are very similar. Curiously, Route Tracing with a mean task duration of about 27 seconds, roughly a quarter the duration of Just Drive, also shows more longitudinal variation than Just Drive. There are two possible longitudinal control behaviors that can be inferred from these results.

The first is that of a short-term "hold" occurring when a subject sheds the car following task to attend to the secondary in-vehicle task. Like steering, if a driver simply holds the accelerator pedal in the same position, for some short period of time the longitudinal position of the car relative to the lead vehicle will not change appreciably. After some period, however, numerous factors including friction, wind drag, and road surface will require an adjustment of the accelerator pedal to maintain longitudinal position. Thus, a short duration task with longitudinal variability significantly less than for Just Drive may be an indicator of a distracted state where a driver is not actively engaged in car following. A study of time series data and examination of accelerator pedal actuation and its relation to the longitudinal metrics may confirm this type of distraction.

The opposite condition, larger longitudinal variations, may be indicative of "falling back". In this situation, which may start as a short-term hold, a driver is not actively attending to car following and begins to fall back from the lead vehicle. This is indicated by high-positive range rates, increased range and decreasing speed. At some point, the driver returns attention to the lead car and accelerates the vehicle to "catch up". This is indicated by higher closing rates (negative range rate), decreased range, and increased speed. A time series study of driver behavior may confirm this condition is occurring by examining accelerator pedal position as well as the longitudinal metrics presented here. Major metrics of such an examination that may be useful would be frequency and amplitude of the variation in longitudinal position.

3.7 Discriminability Analyses

In examining driving performance data, the DWM project identified metrics in each of four categories that were sensitive to intrusion from in-vehicle tasks: lateral control, longitudinal control, event detection, and eyeglance. The objective was to produce a set of driving performance measures against which surrogate measures could be assessed and selected.

Within each category, it was essential that the selected driving performance measures were repeatable, valid, and could make discriminations appropriate for the issues surrounding effects of in-vehicle tasks on driving performance. This was important so that the set of driving performance measures produced from roadway testing (i.e., the test track, since that was the only venue in which all of the 23 DWM tasks could be tested) could serve as a context for selecting and interpreting analyses of surrogate measures.

Prior analyses in this chapter addressed the issue of repeatability for each class of measures, and identified specific metrics within each of the four categories of measurement that were repeatable. In terms of validity, the measures in each of the four categories of measurement were deemed to have "content validity" (Adcock, and Collier, 2001; Carmines and Zeller, 1991) by virtue of their link to highway safety as indicated in Chapter 2. As such, formal analyses of intercorrelations among the measures were not done to assess predictive validity among the measures, with one exception—intercorrelations between eyeglance metrics and selected other driving performance measures. This was done to help provide insights into which of the large number of repeatable eyeglance metrics might be most useful. Therefore, in the discriminability analyses, repeatable measures from the four categories were examined, and their validity was taken as a given, since they were assumed to be relevant measures of driving performance from each of the four categories.

In this section, analyses of discriminability are defined and examined and the results of all analyses are pulled together into a coherent set of driving performance measures against which the subsequent evaluation of surrogate measures can be interpreted.

3.7.1 Fundamental Concepts Underlying Discriminability Analysis

Discriminability analyses in this project were based on two fundamental things:

1. A determination of the types of discrimination that were required, and how fine they might be. What is required is the definition of two categories that are to be discriminated, along with a sorting of tasks into those categories in a manner accepted as consistent with prior prediction based on the literature, theory, modeling from data unrelated to data collected for this project, and engineering judgment.
2. Alignment of metrics with a statistical test that can be used to determine whether two tasks are discriminably different in the predicted direction. What is required is an interpretation of a metric that allows a one-tailed test of discrimination (e.g., a higher score on a metric indicates a higher workload or, alternatively and when appropriate, a lower score on a metric indicates higher workload).

In evaluating whether a metric can make appropriate discriminations to be useful in the product development process, both of these inputs must be carefully considered.

3.7.1.1 Types and Degree of Discrimination Required

Level 1 Discrimination

The first level of discrimination that was important for evaluating whether a new task or device is overloading or intruding on driving performance, was to determine whether a multitasking state, which would consist of doing a subsidiary task while driving, could be discriminated from just driving. If at least these two categories could be discriminated on a given metric, it would be useful insofar as it would indicate that a metric could distinguish that an extra load is being carried by the driver or is intruding upon driving in a measurable way.

It was recognized that Level 1 might be a very easy level of discrimination if multitasking states intruded extensively on driving performance, and these effects were large in magnitude. However, Level 1 discriminations might also be fine discriminations, and difficult to make, if some multitasking states resulted in only subtle levels of interference with driving, and influenced driving in barely detectable ways. The degree of coarseness or fineness in the discrimination might also be specific to the metric, and/or specific to each task.

Level 2 Discrimination

If a metric was capable of at least discriminating the multitasking state from just driving, it was of interest to know whether the metric could be used to make additional discriminations between the set of higher-workload and lower-workload tasks. This would be done separately for visual-manual, auditory-vocal, and Just Drive tasks because of their substantially different nature. This was a second level of discrimination of high practical importance to product development efforts. Level 2 discrimination would be critical to distinguish alternative designs for a function or feature. Level 1 discrimination would also be important to comparisons with a standard other than Just Drive.

It was recognized that Level 2 might be a very difficult level of discrimination when the higher and lower workload tasks within a set or relative workload category (based on prior prediction as explained in Chapter 2) were very similar. Level 2 might also be a very easy level of discrimination when the higher and lower workload tasks within a set or relative workload category were very different. However, when this level of discrimination can be achieved, then the metric has some degree of precision to identify those tasks that are more likely to cause intrusion on driving performance, and less likely to falsely identify tasks that do not significantly intrude on driving. How much precision a metric possesses in making discriminations depends largely on the gradient of difference that exists within the task set undergoing evaluation between lower and higher workload on each metric. A metric's discrimination or sensitivity also depends on measurement error, i.e., a measure's susceptibility to error variance or measurement variation due to unknown or extraneous causes.

It had been hoped that the surrogate metrics selected for the DWM toolkit would achieve the outcome of being able to distinguish lower- and higher-workload tasks within a set. However, it was not possible in advance to guarantee an even spread of tasks along a gradient from low to high workload within each set of tasks (visual-manual and auditory-vocal), and on every metric. It could also be limited by the sensitivity of a given measure to task differences rather than the selection of tasks per se. Therefore, the binary classification into higher-workload and lower-workload tasks, based on prior prediction, was used to frame discriminability analyses. An additional hope was to be able to evaluate those surrogate metrics using driving performance measures, including eyeglance measures, taken directly from driving (either on-the-road or, as in this case, from the test track) that also met both Level 1 and Level 2 discriminability criteria.

3.7.1.2 Application of Discriminability Analyses to Driving Performance Metrics and Surrogates

In the analysis framework set up for this project, discriminability analysis was used twice on the driving performance measures reported in this chapter. It was used first to determine if there was observable intrusion of tasks that could be discriminated from Just Drive on measures of driving performance from the test track that were valid (by virtue of being linked to measures of real world driving performance), and repeatable (based on split-half repeatability analyses) that could be discriminated from just driving. Second, discriminability analysis was used to determine whether those measures that were valid and repeatable could discriminate higher- from lower-workload tasks (within each major type of task, visual-manual and auditory-vocal).

It would seem that in this two-tiered application of discriminability analysis, Level 1 (discrimination of concurrent performance of secondary tasks while driving from driving performance alone) is necessary, and, in some special cases, sufficient. But Level 2 discrimination of higher-workload from lower-workload tasks is of great practical value in product development. Ideally, measurement discrimination or sensitivity would produce

interpretable and statistically significant differences between sets of higher-workload tasks versus lower-workload tasks. Note, however, that low-workload and high-workload tasks may look similar on a metric that meets only Level 1 discriminability. Yet it may be the case that some metrics are important enough, due to the role they play in driving or in traffic situations, that discrimination from Just Drive is of interest or value even though low-workload tasks cannot be discriminated from high-workload tasks. This may occur when all tasks of a certain type similarly affect a metric, for example, an event detection metric. Alternatively, it may be that the gradient of difference between higher- and lower-workload tasks on a metric is sometimes too narrow to permit a discrimination between very tiny differences in workload (as ascertained by other measures or methods), and thus does not justify requiring a Level 2 outcome. These special conditions might warrant acceptance of a Level 1 discriminability outcome. Thus, while Level 1 may be adequate for some special metrics or purposes, Level 2 discriminability is desirable and should confer on metrics a special status as "exceptionally good" whenever it is achieved.

3.7.2 Alignment of Metric Interpretation with One-tailed Statistical Tests

To apply discriminability analysis as it has been defined on the DWM project, it is necessary to test directionally for the predicted outcome. This is necessary because the higher-workload and lower-workload prior predictions produce binary task categories that are ordinal and not simply labels (Nunnally, 1978). As indicated elsewhere, discriminability analysis was performed separately for visual-manual tasks and for auditory-vocal tasks and Just Drive combined. Mixed-mode tasks were not assessed because less is known about their properties and the performance impact of the interactive voice response systems used with both mixed-mode tasks.

Directional hypotheses can take two forms. For example, as workload increases for a visual-manual task, the mean number of glances to task-related areas would also be expected to increase. This is referred to as the "more (i.e., greater magnitude) is more (workload)" prediction alignment (meaning "as the metric scale measured more of the underlying factor, more workload was associated with it."). For other measures, as workload increases for a visual-manual task, the measure is expected to decrease. This is referred to as a "less (i.e., lesser magnitude) is more (workload)" prediction alignment (meaning "as the metric scale measured less of the underlying factor, more workload was associated with it."). For example, as workload increased for a visual-manual task, the mean eyes-on-road time would be expected to decrease, all else being equal). In all cases, the prediction alignments in the other direction ("more is less") were also tested for measures whose content validity justified it.

3.7.2.1 Level 1 Discriminability Analysis

For Level 1 analyses, each task which was done concurrently while driving was tested against the Just Drive task. Thus, the Just Drive task was put on the low-workload side of the matrix, and all other tasks were put on the high-workload side of the matrix. However, low-workload tasks were summarized separately from high-workload tasks. A sample matrix is shown in Figure 3-54.

Task Classifications for Comparisons with Just Drive

	Auditory Vocal Tasks	Visual Manual Tasks
Lower Workload In-Vehicle Tasks	Sports Broadcast Book on Tape Listen Book on Tape Summarize Biographic Q & A	HVAC Radio Easy Radio Hard Cassette CD/Track 7 Coins
Higher Workload In-Vehicle Tasks	Route Instruction Route Orientation Travel Computations	Manual Dial Read Easy Read Hard Map Easy Map Hard Route Tracing Destination Entry

Exploratory (Not Used): Voice Dial
Delta Flightline

Figure 3-54. Task Classifications Used for Level 1 Discriminability Analyses

Note: In these analyses, tasks in this matrix were all compared to Just Drive as a lower-workload task since there was no subsidiary task at all.

3.7.2.2 Level 2 Discriminability Analysis

For Level 2 analyses, tasks were sorted into high- and low-workload categories, based on a combination of findings from the literature, analytical modeling, and expert judgment (See Chapter (2). These sorts are shown in Figure 3-55.

Task Classifications for Comparisons of Low and High Workload

	Auditory Vocal Tasks & Just Drive	Visual Manual Tasks
Lower Workload In-Vehicle Tasks	Sports Broadcast Book on Tape Listen Book on Tape Summarize Biographic Q & A Just Drive	HVAC Radio Easy Radio Hard Cassette CD/Track 7 Coins
Higher Workload In-Vehicle Tasks	Route Instruction Route Orientation Travel Computations	Manual Dial Read Easy Read Hard Map Easy Map Hard Route Tracing Destination Entry

Exploratory (Not Used): Voice Dial
Delta Flightline

Figure 3-55. Task Classification for Level 2 Discriminability Analyses

Note: In these analyses, tasks shown as high were compared to tasks shown as low within each type of task. The Just Drive task was grouped as a lower-workload task within the auditory-vocal task set, due to the similarity of its length and performance profiles to these tasks. Like auditory-vocal tasks, Just Drive had no visual input demands from a subsidiary task or manual output demands from a subsidiary task.

If a task was judged to fall within the higher-workload category group than the comparison task, it was predicted that each subject's pairs of task scores for these tasks, when compared, would result in a difference score that was positive (+), matching the predicted alignment (Task 1 greater than Task 2, on a given a metric, if that metric followed the expected "higher on the metric means higher-workload" pattern or Task 1 less than Task 2 for a "lower on the metric means higher-workload" pattern.). This directional hypothesis was then tested using a sign test. The sign test was applied at the per-participant level for each selected measure. For each participant, the difference was calculated between that person's performance measure on one task (e.g., Task A) versus that same person's performance measure on another task (e.g., Task B) under comparison. Only the sign of each difference per participant (+ if Task A > Task B; if Task A < Task B; or tie if Task A = Task B) was retained for analysis. The signs of the differences were tallied across all participants who performed both of the tasks (e.g., Task A and Task B) under comparison. The distribution of positive and negative signs (ignoring ties) was then evaluated statistically. The evaluation assessed whether the percentage of positive signs and negative signs was much different from 50-50 by chance. A directional sign test specifically looked to see whether or not the percentage was significantly different from 50-50 and in the expected direction. This method was used for both Level 1 and for Level 2 discriminability tests. The sign test examined ordinal differences between a person's performance on one task and another task. This is a very different level of analysis than group means, medians, percent misses, and the like. The test's main virtues

are its long history, simplicity, and its freedom from assumptions of normality, linearity, homogeneity of variance, and so forth.

The discriminability percentages were based on varying numbers of paired-comparisons. (Refer to Figure 3-54 for the breakout of CAMP DWM tasks into higher- and lower-workload categories based on prior prediction). The number of paired comparisons depended on whether Level 1 results were reported or Level 2 results were reported. The following numbers apply for Level 1 comparisons of visual-manual tasks to Just Drive. Visual-manual task discriminability percentages were based on up to 13 comparisons to Just Drive if all 13 visual-manual tasks were included. There were seven paired comparisons to Just Drive if only the seven higher-workload visual-manual tasks were included. There were six paired comparisons to Just Drive if only the six lower-workload visual-manual tasks were included. The following numbers apply for Level 1 comparisons of auditory-vocal tasks to Just Drive. Auditory-vocal task discriminability percentages were based on up to seven paired comparisons if all auditory-vocal tasks were included. There were three paired comparisons to Just Drive if only the three higher-workload tasks were assessed. There were five paired comparisons to Just Drive is only the five lower-workload auditory-vocal tasks were included. These numbers should be kept in mind when considering the Level 1 results.

Level 2 discriminability percentages were based on larger numbers of paired-comparisons. The visual-manual tasks provided seven higher-workload tasks, each compared against six lower-workload tasks. Thus, Level 2 visual-manual task discriminability percentages were based on 42 paired comparisons. The auditory-vocal and Just Drive tasks provided three higher-workload tasks, each compared against five lower-workload tasks. Level 2 auditory-vocal task discriminability percentages were therefore based on 15 paired comparisons. These numbers should also be kept in mind when considering the discriminability results.

3.7.3 Discriminability Results: Levels 1 and 2

Results of discriminability analyses for both Levels 1 and 2 are summarized in Table 3-13, Table 3-14, Table 3-15, and Table 3-16. Table 3-13 shows the results for driving performance metrics for visual-manual tasks plus a small subset of eyeglance measures to the not road location. For these measures a higher score on the metric was believed to indicate higher workload. Table 3-14 shows results for driving performance metrics for auditory-vocal tasks plus these few eyeglance measures. Table 3-15 shows results for eyeglance metrics for visual-manual tasks. Table 3-16 shows results for eyeglance metrics for auditory-vocal tasks. Percentages over 67 percent were highlighted in yellow, but 70 percent was used as a cutoff point for meeting the discriminability criterion.

Level 1 discriminability analyses are summarized for low-workload tasks versus Just Drive in the first column of Table 3-13 and Table 3-14, and are summarized for high-workload tasks versus Just Drive in the second column of Table 3-13 and Table 3-14. The third column summarizes “all tasks” versus Just Drive.

Level 2 discriminability analyses are summarized for high-workload tasks versus low-workload tasks in the fourth column of Table 3-13 and Table 3-14 for driving performance metrics.

Table 3-13. Summary of Level 1 and 2 Discriminability Results for Driving Performance Metrics Based on Test Track Data for Visual-Manual Tasks

Driving Performance Metrics Discriminability Summary
Based on Track Data and Sign Test

	Visual-Manual Tasks			
Metric	Low Workload In-Vehicle Tasks vs. Just Drive	High Workload In-VehicleTasks vs. Just Drive	All In-Vehicle Tasks vs. Just Drive	Low Workload Tasks vs. High Workload Tasks
Lanex Cross Trial	0%	29%	16%	29%
MeanduratNR	0%	29%	16%	90%
MeanglancesNR	0%	14%	8%	81%
MeanmeanNRdur	83%	100%	92%	50%
MeanpctdurNR	100%	100%	100%	45%
PctMissCHMSL	83%	86%	85%	14%
PctMissDecel	100% ▲	57%	77%	2%
PctMissFVTS	100%	100%	100%	2%
SDLP	0%	43%	23%	74%
Speed Diff	0%	14%	8%	88%
Task Duration	0%	0%	0%	90%

▲ Note: A portion of LVD events were not detectable within the task length of short (low-workload) visual-manual tasks. The LVD event was problematic for shorter tasks because of reasons that will be discussed in Chapter 8. It was not possible to correct individual percent missed detections based on the detection threshold. Therefore, the discriminability value of 100 percent shown in the table should be viewed as an overestimate.

3.7.3.1 Level 1 Discriminability Results: Discriminating Visual-manual Tasks from Just Drive using Driving Performance Measures

As is apparent in Table 3-13, in the Level 1 analysis (distinct from the Level 2 analysis) concurrent performance of both low- and high-workload visual-manual tasks during driving were discriminable from Just Drive on two types of measures:

- **Event-detection measures** (percent missed detections for CHMSLs, FVTS events, and LVD events, although LVD percent missed detections must be treated with caution), and
- **Eyeglance measures** (mean duration of all glances to areas other than the roadway, and mean percent of task duration spent looking at non-roadway locations).

However, on measures of lanekeeping and speed keeping, concurrent performance of visual-manual tasks could not be discriminated from just driving. This same set of results held when results were summarized across all visual-manual tasks. These are important findings relative to the selection of surrogate metrics and methodological issues with tasks of shorter duration compared to tasks of longer duration.

It should be noted that discriminability analyses were also conducted on three driving performance measures using a "more is less workload" alignment (SDLP, Speed Difference, and Task Duration). These analyses indicated that visual-manual tasks were discriminable from Just

Drive on the measure of Speed Difference, with a discriminability score of 85 percent (and also on Task Duration, with a discriminability score of 100 percent—the latter of which is not surprising, since Just Drive is a two minute task, and the visual-manual tasks are all considerably shorter). The fact that visual-manual tasks were discriminable on Speed Difference (insofar as they exhibited less speed difference, on average, than Just Drive) is due in part to their shorter duration (which was also discriminable from Just Drive), but also perhaps to shedding of active speed control during a short task (e.g., through hypothesized temporary holds on the accelerator pedal, for instance, as discussed earlier in this chapter).

3.7.3.2 Level 2 Discriminability Results: Discriminating within the Visual-Manual Category – Higher-Workload from Lower-Workload Visual-Manual Tasks

The fifth column of Table 3-13 contains the Level 2 discriminability results for visual-manual tasks. The table indicates higher-workload and lower-workload tasks were acceptably discriminated by the lateral control measure of Standard Deviation of Lane Position (SDLP). They were discriminated well by the longitudinal control measure of Speed Difference (SpeedDiff). Visual-manual tasks were discriminated consistent with prior prediction on selected not-road (NR) eyeglance measures. As will be seen in Table 3-14, visual-manual tasks were also well discriminated by other eyeglance measure, such as task-related eyeglance measures. Finally, these tasks were well discriminated by task duration. This last measure is important because visual-manual tasks had task-intrinsic durations, i.e., durations intrinsic to the task and not of arbitrarily fixed lengths. These measures had task duration as a common component. The implication of this for tasks with task-intrinsic durations will be discussed in Chapter 8.

Visual-manual tasks did not discriminate well between higher-workload and lower-workload tasks on any measure of object and event detection. This result is also related to task duration in a paradoxical way. It too will be treated in the discussion sections of Chapter 8 of this report.

Table 3-14. Summary of Level 1 and 2 Discriminability Results for Driving Performance Metrics Based on Test Track Data for Auditory-Vocal Tasks

Driving Performance Metrics Discriminability Summary
Based on Track Data and Sign Test

	Auditory-Vocal Tasks			
Metric	Low Workload In-Vehicle Tasks vs. Just Drive	High Workload In-VehicleTasks vs. Just Drive	All In-Vehicle Tasks vs. Just Drive	Low Workload Tasks vs. High Workload Tasks
Lanex Cross Trial	0%	0%	0%	0%
MeanduratNR	0%	0%	0%	20%
MeanglancesNR	0%	0%	0%	20%
MeanmeanNRdur	0%	0%	0%	20%
MeanpctdurNR	0%	0%	0%	20%
PctMissCHMSL	25%	0%	14%	0%
PctMissDecel	0%	0%	0%	7%
PctMissFVTS	25%	100%	57%	33%
SDLP	0%	0%	0%	20%
Speed Diff	25%	67%	43%	40%

3.7.3.3 Discriminating Auditory-Vocal Tasks from Just Drive Using Driving Performance Measures

The results in Table 3-14 for the Level 1 analysis (columns 1, 2 and (3) indicate that low-workload auditory-vocal tasks were not discriminable from Just Drive on the driving performance measures in this table. But the high-workload auditory-vocal tasks were discriminable on only one measure (percent missed detections for FVTS). It is interesting to note, that the percent missed detections for FVTS events was still lower than for visual-manual tasks. Speed Difference, though discriminable at 67 percent on the three highest workload tasks only, could not meet the criterion because only three paired comparisons were made. Thus, if one of the three auditory-vocal tasks could not be discriminated from Just Drive, the discriminability score would be 67 percent. Examining the actual values for Speed Difference, auditory-vocal tasks fell in the range between ~5.5 to 8 ft/sec (whereas Just Drive was ~7 ft/sec). Overall for the set of auditory-vocal tasks as a whole, none of the driving performance metrics in the table permitted them to be discriminated from Just Drive.

3.7.3.4 Discriminating within the Auditory-Vocal Category: High- from Low-Workload Tasks

Results in Table 3-14 for the Level 2 analysis (column 4) indicate that high- versus low-workload auditory-vocal tasks could not be discriminated from each other on any of the driving performance measures summarized in this table.

Table 3-15. Summary of Level 1 and 2 Discriminability Results for Eyeglance Metrics for Visual-Manual Tasks, based on Test Track Data

Eye Glance Metrics Discriminability Summary
Based on Visual-Manual Tasks, Test Track Data and Sign Test

	Visual-Manual Tasks							
	More is More Workload				More is Less Workload			
Eyeglance Metric	Low Workload In-Vehicle Tasks vs. Just Drive	High Workload In-VehicleTasks vs. Just Drive	All In-Vehicle Tasks vs. Just Drive	Low Workload Tasks vs. High Workload Tasks	Low Workload In-Vehicle Tasks vs. Just Drive	High Workload In-VehicleTasks vs. Just Drive	All In-Vehicle Tasks vs. Just Drive	Low Workload Tasks vs. High Workload Tasks
MaxRDdur	0%	0%	0%	38%	100%	100%	100%	17%
MaxTdur	0%	0%	0%	36%	100%	100%	100%	5%
MaxTRdur	0%	0%	0%	43%	0%	0%	0%	2%
MeanduratMR	0%	0%	0%	60%	100%	100%	100%	0%
MeanduratRD	0%	0%	0%	55%	100%	100%	100%	2%
MeanduratSA	0%	0%	0%	50%	100%	100%	100%	2%
MeandurateTR	0%	0%	0%	74%	0%	0%	0%	0%
MeanglncesMR	0%	0%	0%	62%	100%	100%	100%	0%
MeanglncesRD	0%	14%	8%	76%	100%	86%	92%	2%
MeanglncesSA	0%	0%	0%	52%	100%	100%	100%	0%
MeanglncesTR	0%	0%	0%	79%	0%	0%	0%	2%
MeangrateRD	100%	100%	100%	7%	0%	0%	0%	45%
MeangrateTR	0%	0%	0%	21%	0%	0%	0%	19%
MeanmeanMRdr	0%	0%	0%	2%	100%	100%	100%	5%
MeanmeanRDdr	0%	0%	0%	26%	100%	100%	100%	38%
MeanmeanTRdr	0%	0%	0%	0%				
MeanmeanSAdr	0%	0%	0%	5%	83%	100%	92%	12%
MeanmedRDdur	0%	0%	0%	10%	100%	100%	100%	38%
MeanmedSAdur	0%	0%	0%	2%	83%	100%	92%	2%
MeanpctdurRD	0%	0%	0%	7%	100%	100%	100%	43%
MeanpctdurTR	0%	0%	0%	40%	0%	0%	0%	12%
MeansdRDdur	0%	0%	0%	19%	100%	100%	100%	36%
MeansdTdur	0%	0%	0%	36%	100%	100%	100%	7%
MeansdTRdur	0%	0%	0%	36%	0%	0%	0%	0%
MeanTaskdur	0%	0%	0%	81%	100%	100%	100%	0%
MeanTglsprs	100%	100%	100%	12%	0%	0%	0%	31%
MeanTskglncs	0%	14%	8%	81%	100%	86%	92%	2%
TotalTRdur	0%	0%	0%	74%	0%	0%	0%	0%

3.7.3.5 Discriminating Visual-Manual Tasks from Just Drive Using Eyeglance Metrics

In terms of the broader set of eyeglance metrics, Level 1 and 2 results are shown in Table 3-15. The “more is more workload” alignment of metric scale with hypothesized workload effects is shown on the left side of the table, and the “more is less workload” alignment is shown on the right side of the table (meaning that a higher score on the metric would indicate a lower-level of workload and a lower score on the metric would indicate a higher level of workload). A good example of this latter alignment would be mirror scanning: glances to the mirrors (MeanglancesSA and MeanglancesMR) decreased as workload increased.

As is clear from Table 3-15, (columns 1-3, on the left side) for visual-manual tasks, only two eyeglance metrics enabled discrimination of visual-manual tasks from Just Drive using the “more is more workload” alignment of metric and hypothesis. These were the metrics discriminating visual-manual tasks from Just Drive (“more is more” alignment):

MeangrateRD **mean glance rate to the roadway location**

MeanTglsprs **mean total glance rate per second**, or the rate of glances per second, considering all glances to all locations during the task

Glance rate increased for visual-manual tasks, due to the fact that drivers looked back and forth between the in-vehicle task and the roadway. This increased the number of glances to the road and to all locations in total. This means that the glances were short (under 2 seconds, on average) or that the glance rates were greater for shorter duration tasks, or both. This pattern is discriminable from Just Drive, based on the analysis of glances to the road during Just Drive and the univariate analyses, discussed previously in this chapter. The glances for Just Drive were fewer in number, longer (8 seconds, on average), and thus slower in rate-per-second. Duration differences may account for the failure of "more of a measure is more workload" comparisons. Recall that the Just Drive task (a two-minute task) was much longer than visual-manual tasks (generally under 20 seconds, typically, except for Destination Entry).

On the right side of the table, columns 1 through 3 indicate that a great many eyeglance metrics discriminated high-workload visual-manual tasks from low-workload tasks, using the “more is less” alignment of these metrics with workload. Not all of the alignments tested in this way are meaningful. Those that are, however, included discriminably shorter glances to the road and mirrors/SA areas, discriminably fewer glances to the road, discriminably fewer glances to road and mirrors, and discriminably lower percent of task-time spent viewing road and mirrors. The metrics related to task-related glancing dropped out as discriminating visual-manual tasks from Just Drive (because there are no task-related glances in Just Drive).

3.7.3.6 Discriminating within the Visual-Manual Category: High- from Low-Workload Tasks

The fourth column on the left side of the table ("more is more workload alignment") indicates that four eyeglance metrics enabled discrimination of high-workload visual-manual tasks from low-workload visual-manual tasks. These were metrics discriminating high- from low-workload visual-manual tasks ("more is more" alignment):

Mean durat TR — **mean total glance time spent on the in-vehicle task.** Indicated that high-workload visual-manual tasks had discriminably longer total glance times to task-related areas than did the low-workload visual-manual tasks.

Total TR dur — **another way to derive Mean durat TR, the prior entry in this list**

Indicated that high-workload visual-manual tasks had discriminably longer total glance times to task-related areas than did the low workload visual-manual tasks.

MeanTaskdur — **mean task duration computed from eyeglance data**

Indicated that high-workload visual-manual tasks had discriminably longer task durations (time to complete the task; based on eye data) than did the low-workload visual-manual tasks.

Mean glances RD — **mean number of glances to the road**

High-workload visual-manual tasks had discriminably more glances to the road than did low-workload visual-manual tasks. (This is because a driver looks back and-forth between task and road to perform the visual-manual task, increasing both number of glances to road and task. As the task lengthened, the number of glances to both areas accumulated.)

MeanglancesTR — **mean number of glances to task-related areas**

High-workload visual-manual tasks had discriminably more glances to the task than did low-workload visual-manual tasks (per the discussion above).

Mean Taskglances — **mean number of all glances occurring during task**

High-workload visual-manual tasks had discriminably more glances to all areas during the task than did low-workload tasks.

The fourth column on the right side of the table indicates that none of the eyeglance metrics enabled discrimination of high- from low-workload visual-manual tasks, when used in the "more is less workload" alignment.

Table 3-16. Summary of Level 1 and 2 Discriminability Results for Eyeglance Metrics for Auditory-Vocal Tasks Based on Test Track Data

Eye Glance Metrics Discriminability Summary Based on Auditory-Vocal Tasks, Test Track Data and Sign Test								
	Auditory-Vocal Tasks							
	More is More Workload				More is Less Workload			
Eyeglance Metric	Low Workload In-Vehicle Tasks vs. Just Drive	High Workload In-VehicleTasks vs. Just Drive	All In-Vehicle Tasks vs. Just Drive	Low Workload Tasks vs. High Workload Tasks	Low Workload In-Vehicle Tasks vs. Just Drive	High Workload In-VehicleTasks vs. Just Drive	All In-Vehicle Tasks vs. Just Drive	Low Workload Tasks vs. High Workload Tasks
MaxRDdur	25%	67%	43%	33%	0%	0%	0%	0%
MaxTdur	25%	67%	43%	33%	0%	0%	0%	0%
MaxTRdur	0%	0%	0%	0%	0%	0%	0%	0%
MeanduratMR	0%	0%	0%	20%	50%	100%	71%	53%
MeanduratRD	50%	33%	43%	33%	25%	0%	14%	27%
MeanduratSA	0%	0%	0%	20%	50%	100%	71%	53%
MeandurateTR	0%	0%	0%	0%	0%	0%	0%	0%
MeanglncesMR	0%	0%	0%	20%	50%	100%	71%	53%
MeanglncesRD	0%	0%	0%	20%	50%	100%	71%	40%
MeanglncesSA	0%	0%	0%	20%	75%	100%	86%	53%
MeanglncesTR	0%	0%	0%	0%	0%	0%	0%	0%
MeangrateRD	0%	0%	0%	0%	75%	100%	86%	27%
MeangrateTR	0%	0%	0%	0%	0%	0%	0%	0%
MeanmeanMRdr	0%	0%	0%	20%	75%	67%	72%	13%
MeanmeanRDdr	75%	100%	86%	27%	0%	0%	0%	0%
MeanmeanTRdr	0%	0%	0%	0%				
MeanmeanSAdr	0%	0%	0%	20%	75%	67%	72%	20%
MeanmedRDdur	75%	100%	86%	20%	0%	0%	0%	13%
MeanmedSAdur	0%	0%	0%	20%	25%	33%	28%	7%
MeanpctdurRD	75%	100%	86%	27%	0%	0%	0%	13%
MeanpctdurTR	0%	0%	0%	0%	0%	0%	0%	0%
MeansdRDdur	25%	100%	57%	40%	0%	0%	0%	7%
MeansdTdur	25%	100%	57%	47%	0%	0%	0%	0%
MeansdTRdur	0%	0%	0%	0%	0%	0%	0%	0%
MeanTaskdur	50%	33%	43%	33%	50%	67%	57%	47%
MeanTglsprs	0%	0%	0%	0%	75%	100%	86%	27%
MeanTskglncs	0%	0%	0%	20%	50%	100%	71%	40%
TotalTRdur	0%	0%	0%	0%	0%	0%	0%	0%

3.7.3.7 Discriminating Auditory-Vocal Tasks from Just Drive Using Eyeglance Metrics

As is clear from Table 3-16, (columns 1-3, on the left side of the table) for auditory-vocal tasks, only three eyeglance metrics (two of them closely related) allowed auditory-vocal tasks to be discriminated from Just Drive using the "more is more" alignment of metric to workload hypothesis. The three measures that discriminated both low- and high-workload auditory-vocal tasks from Just Drive (eyeglance metrics discriminating auditory-vocal tasks from Just Drive ("more is more") were:

MeanmeanRDdr **Mean duration of glances to the roadway**

The mean duration of glances to the roadway were discriminably longer for auditory-vocal tasks than for Just Drive.

MeanmedRDdur **Median duration of glances to the roadway**

The median duration of glances to the roadway were discriminably longer for auditory-vocal tasks than for Just Drive.

MeanpctdurRD **Mean percent of task spent looking at the roadway**

The percent of time during the task that was spent looking at the roadway was discriminably larger for auditory-vocal tasks than for Just Drive.

From columns 1-3 on the right side of Table 3-16, it is apparent that a few additional eyeglance metrics permitted discrimination of auditory-vocal tasks from Just Drive using the "more is less" alignment of metric to workload hypothesis. Those eyeglance metrics that discriminated both low- and high-workload auditory-vocal tasks from Just Drive ("more is less") included:

MeangrateRD **Mean Glance Rate to the Road**

Auditory-vocal tasks showed a lower glance rate to the road (due to longer duration roadway glances) that permitted this Glance Rate To The Road variable to discriminate them from Just Drive.

MeanMRdr **Mean Duration of Glances to the Mirror**

Auditory-vocal tasks showed shorter glance durations on mirrors (a consequence of more time spent viewing the road) and this was discriminable from the length of glances to the mirrors for Just Drive.

MeanSAdr **Mean Duration of Glances to the Situation Awareness Areas – Mirrors & Speedometer**

Auditory-vocal tasks showed shorter glance durations on mirrors (a consequence of more time spent viewing the road) and this was discriminable from the length of glances to the mirrors/SA areas for Just Drive.

MeanTglsprs (Mean Total Glance Rate Per Second) to any and all locations

They were also four additional variables that discriminated only high-workload auditory-vocal tasks from Just Drive using the "more is more" alignment:

MaxRDdur **Duration of longest glance to roadway**

Longest roadway glance for high-workload auditory-vocal tasks was discriminable from that for Just Drive.

MaxTdur **Duration of longest glance of any type to any location occurring during task**

Longest glance of any type for high-workload auditory-vocal tasks was discriminable from that for Just Drive.

MeansdRDdur **Mean standard deviation of glances to roadway**

Standard deviation for durations of roadway glances was larger for high workload auditory-vocal tasks and discriminable from that for Just Drive.

MeansdTdur **Mean standard deviation of all glance durations to any location during a task**

Average standard deviation for all glance durations was larger for high workload auditory-vocal tasks and discriminable from that for Just Drive.

Those eyeglance metrics that discriminated only high-workload auditory-vocal tasks from Just Drive using the "more is less" alignment included:

MeanduratMR **(Mean Total Glance Time on the Mirrors)**

MeanduratSA **(Mean Total Glance Time on the Situation Awareness Areas – Mirrors/Speedometer)**

Auditory-vocal tasks showed less total glance time on mirrors (a consequence of more time spent viewing the road), and this was discriminable from the total glance time on the mirrors/SA areas for Just Drive.

MeanglancesMR **(Mean Number of Glances to the Mirrors)**

MeanglancesSA **(Mean Number of Glances to the Mirrors)**

Auditory-vocal tasks showed fewer glances to the mirrors (a consequence of more time spent viewing the road), and this was discriminable from the number of glances to the mirrors/SA areas for Just Drive.

MeanmeanMRdr (**Mean Duration of Glances to the Mirror)**

MeanmeanSAdr **(Mean Duration of Glances to the Mirror)**

Auditory-vocal tasks showed shorter glance durations on mirrors (a consequence of more time spent viewing the road) and this was discriminable from the duration of glances to the mirrors/SA areas for Just Drive.

MeanglancesRD **(Mean Number of Glances to the Road)**

Auditory-vocal tasks showed fewer glances to the road (due to longer duration glances), and this was discriminable from the number of glance to the road for Just Drive.

MeangrateRD **(Mean Glance Rate to the Road)**

Auditory-vocal tasks showed lower glance rates to the road, and this was discriminable from those for Just Drive.

MeanTglsprs **(Mean Total Glance Rate per Second)**

Due to the long gazing at the roadway, auditory-vocal tasks had lower glance rates per second (for total glances during the task) than did Just Drive.

MeanTaskglances **(Mean Number of Total Glances to All Locations During Task)**

The number of glances of all types to all locations was fewer for auditory-vocal tasks than for Just Drive. Again, this was due to fewer, long gazes at the forward roadway.

MeanTaskDur **(Mean Task Duration Based on Eye Movement Data)**

If this alignment of "less is more" is meaningful for this variable, it means that shorter task durations for auditory-vocal tasks (based on the eye movement record) were discriminable from the Just Drive task which was nominally of similar length (though one auditory-vocal task was shorter than Just Drive). It is not clear, however, if this alignment is meaningful for this variable.

3.7.3.8 Discriminating Within the Auditory-Vocal Category: High- From Low-Workload Tasks

The fourth column on the left side of the table (using the "more is more workload alignment") indicates that none of the eyeglance metrics permitted discrimination of high- from low-workload auditory-vocal tasks. The fourth column on the left side of the table (using the "more is less" alignment) similarly shows that none of the eyeglance metrics discriminated on this alignment either, so no Level 2 discriminability was achieved on any eyeglance metric for auditory-vocal tasks.

3.7.4 Summary of Level 1 Discriminability Results

3.7.4.1 Auditory-Vocal Task Results

The results show that auditory-vocal tasks were discriminable from Just Drive, suggesting that there is some subtle intrusion on driving performance. The intrusion is seen predominantly on eyeglance measures related to a concentration of long gazes on the forward roadway. Vehicle control metrics did not show evidence of intrusion from the in-vehicle tasks. Percent Missed Detections of FVTS discriminated the three "higher-workload" auditory-vocal tasks from Just Drive. These findings together (the prolonged gazing at the forward roadway, reduced scanning and durations of glances at the mirrors, and higher percent missed detections for FVTS Events, and elevated Speed Difference) could perhaps be consistent with cognitive loading. The implication for the surrogate toolkit is that methods will be needed for evaluating auditory-vocal tasks so that performance characteristics of these types can be predicted to distinguish such tasks from a Just Drive baseline.

3.7.4.2 Visual-Manual Task Results

In terms of visual-manual tasks relative to Just Drive, the analysis reveals intrusion on driving performance in the areas of eyeglance (MeanmeanNRdur, MeanpctdurNR, each at 100 percent) and event detection (percent missed detections for CHMSL at 85 percent and percent missed detections for FVTS at 100%). Decrements in event detection were larger for visual-manual tasks than for auditory-vocal tasks. Furthermore, the eyeglance analysis of additional measures revealed a number of additional eyeglance metrics that discriminated visual-manual tasks from Just Drive, indicating intrusion of visual-manual tasks on driving performance. The implications

for the surrogate toolkit are that surrogates methods/metrics will be needed to address both of these areas of performance-effect from task workload—visual demand and event detection.

Level 1 results for both types of tasks are summarized, at a high level, in Figure 3-56. This figure should not be used in place of the more detailed findings described above. It is provided only as a summarizing framework. The bracketed items marked with a plus sign (+) indicate metrics that discriminate with a "more is more" alignment. Those marked with a negative sign (–) indicate metrics that discriminate with a "more is less" alignment. Metrics have been grouped together for convenience and meaningfulness in summarizing. In a few instances, metrics that only discriminated high-workload tasks from Just Drive have been included due to the fact that findings on them tended to cluster with the other metrics called out.

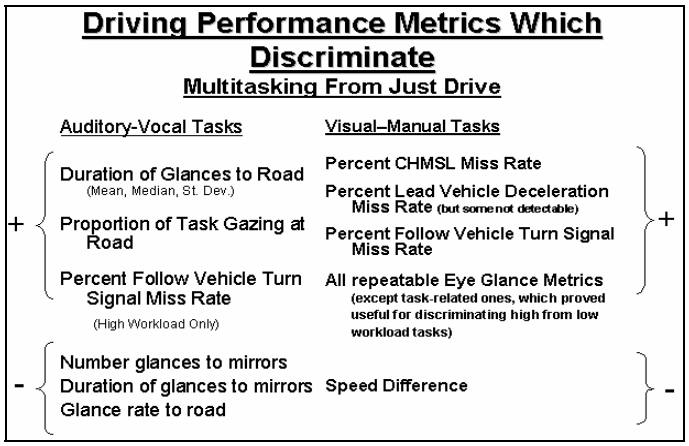


Figure 3-56. High-level Summary of Metrics Found to Discriminate Multitasking From Just Drive

Note: In some instances, metrics have been grouped together where convenient or meaningful.

3.7.5 Summary of Level 2 Discriminability Findings

3.7.5.1 Auditory-Vocal Task Results

No discrimination between high- and low-workload auditory-vocal tasks was possible with any of the performance measures of any type. This may have been because the tasks were too similar (all but one of them were ~2 minutes in length).

3.7.5.2 Visual-Manual Task Results

High- and low-workload visual-manual tasks were discriminable from each other on a number of metrics. These included:

- Task Duration
- SDLP
- Speed Difference
- Glance Metrics, including (but not limited to):
  - Number of glances to the task
  - Total glance time to the task
  - Number of all glances occurring during task

The implication of this for the surrogate development effort is that, in addition to measures of event detection, surrogates for these driving performance metrics may be important to include.

A summary of the Level 2 Discriminability analysis appears in Figure 3-57 for both task types.

As mentioned, surrogates will need to cover these areas of performance-effect in addition to those emerging from the Level 1 analysis.

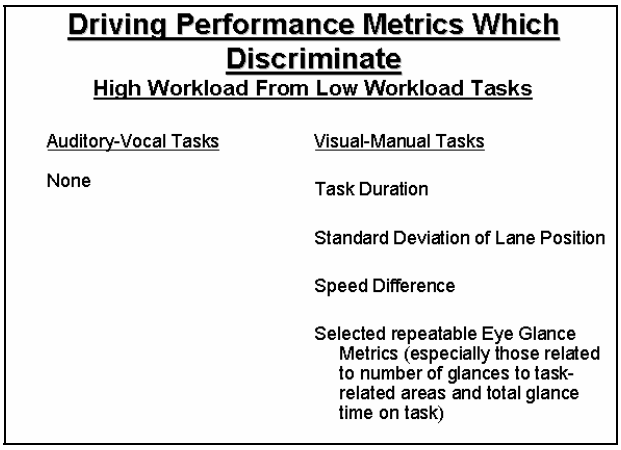


Figure 3-57. High-Level Summary of Metrics Found to Discriminate High From Low Workload Within Each Type of Task

3.7.6 Visualization of the Multidimensional Performance Effects

In approaching the development and selection of surrogate metrics in Chapter 5, it is essential to have a firm understanding of the performance profiles that emerged from the test track data. Toward that end, the star charts shown in Figure 3-58 were created. In those charts, there are 10 radials, each depicting a different driving performance metric: SDLP (standard deviation of lane-keeping), Percent Trials with a Cross of the Lane Line, Speed Difference, LVD Percent Missed Detections, CHMSL Percent Missed Detections, FVTS Percent Missed Detections, Duration of Road Glances, Number of Task Glances, # Mirror Glances, and Total Glance Rate. These are all variables that were repeatable and in some way emerged from the discriminability analysis. The data on each metric was standardized (by taking the matrix of task-level means on each metric, and converting them to standard scores or z-scores). Across each set of z-scores on a metric, the mean of the tasks will be zero and the standard deviation will be one. However, the shape of the original distribution (e.g., skewness, extreme values, etc.) remains the same. These standard scores allow the metrics to be plotted on the same dimensionless scale, and enables comparisons to be made of the size and shape of the "stars" on each plot. One plot in Figure 3-58 shows the Just Drive task, one shows the average visual-manual task (obtained by taking the mean of all visual-manual tasks on each metric, and converting that to a z-score), and one shows the average auditory-vocal task.

By comparing the stars, it becomes apparent that the visual-manual tasks had a more pronounced effect on event detection (elevating percent missed detections above those for Just Drive or auditory-vocal tasks), a pronounced effect on glance durations to the road and number of glances to the mirrors (reducing them relative to Just Drive and auditory-vocal tasks), an effect on number of task glances and total glance rate (increasing them). However, on the measures of SDLP and Speed Difference, performance was less than (not greater than) Just Drive, and on Percent Lane Cross Trials, it was fairly comparable, on average (it was high only for the three highest visual

manual tasks, and that effect is less apparent due to averaging in the star chart). In a similar way, the auditory-vocal star chart can be compared to the others. It is easy to see the effect on FVTS percent missed detections (higher than Just Drive but lower than visual-manual tasks), and lengthened duration of roadway glances. Performance on SDLP, Percent Cross Trials, and Speed Difference has a similar appearance to Just Drive. These charts again emphasize the importance of ensuring that the surrogate methods and metrics cover event detection and visual demand.

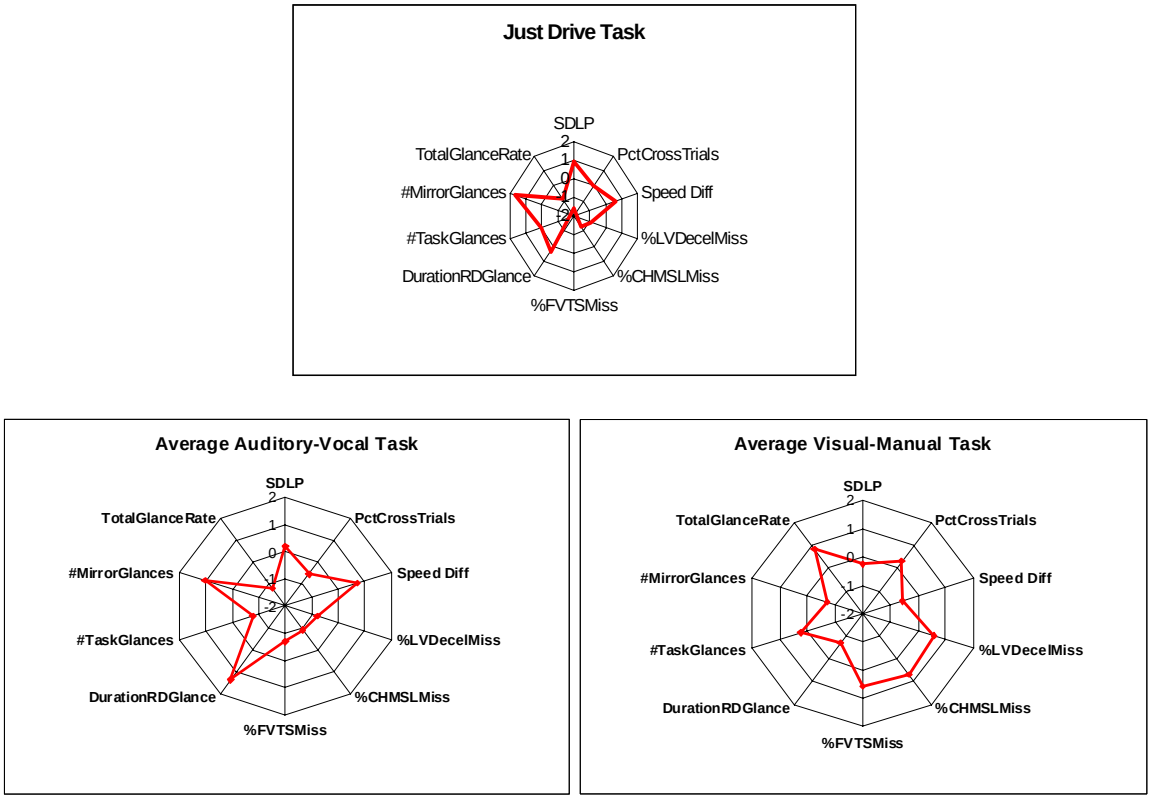


Figure 3-58. Star Charts of Just Drive, the Average Visual-Manual, and Average Auditory-Vocal Task

Note: Points displayed on radials represent standardized scores where the mean of all tasks equals zero.

3.7.7 Closing Comments

The discriminability results integrate the findings from driving performance metrics into a coherent picture for each type of task. In addition, they identify driving performance measures against which the evaluation of surrogate measures will be most productive and indicate for which areas of performance surrogate methods and metrics are needed.

3.8 Chapter References

Adcock, R., and Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. American Political Science Review, 95(3), 529-546.

Allen, R. W., Parseghian, Z., and Stein, A. C. (1996). A driving simulator study of the performance effects of low blood alcohol concentration. Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting, 943-946.

Brown, I. D. (1994). Driver fatigue. Human Factors, 36(2), 298-314.

Carmines, E., and Zeller, R. (1991). Reliability and validity assessment. London: Sage Publications.

McLean, J. R., and Hoffman, E. R. (1975). Steering reversals as a measure of driver performance and steering task difficulty. Human Factors, 17(3), 248-256.

Nunnally, J. (1978). Psychometric theory (2nd Edition). New York: McGraw-Hill.

Verbeke, G. and Molenberghs, G. (1997), Linear Mixed Models in Practice: A SAS-Oriented Approach, New York: Springer-Verlag.

Young, R. A., and Angell, L. S. (2003). The Dimensions of Driver Performance During Secondary Manual Tasks. Proceedings of “Driving Assessment 2003: The Second International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design.” Park City, Utah.

4 On-Road Results

4.1 Background

The rationale and motivation for the driving performance categories and measures selected in the CAMP DWM project are discussed in Chapter 1, *Introduction* and Chapter 3, *Test Track Results*.

Compared to the track, the road venue involved testing a smaller set of visual-manual tasks. The tasks reserved for track testing were those that, in the judgment of the research team, were best done on a closed course because of their expected higher workload. Thus, it is important to keep in mind that the road venue included a smaller set of visual-manual tasks but the same auditory-vocal tasks.

Summary statistics for all measures reported in this chapter are provided in Appendix Q.

4.2 Participants

One hundred one licensed drivers were recruited from the Detroit metropolitan area for participation in the on-road phase of the study. The same screening procedures were used for the on-road study as were used for the track study discussed in Chapter 3. Table 4-1 presents the distribution of the age and gender of the participants. The participants were paid $320 for their two-day time commitment to the study. Additional details about the sample of participants and the screening process can be found in the appendices to this report.

Table 4-1. Age and Gender of On-Road Participants

	Age Category						
	20's	30's	40's	50's	60's	70's	All
Male	9	7	9	7	8	9	49
Female	8	10	7	12	10	5	52
All	17	17	16	19	18	14	101

4.3 Road Task Effects on Object-and-Event Detection (OED)

To determine if roadway events were monitored, the participants were presented with events to be detected as part of on-road and test-track driving.

4.3.1 Center High-Mount Stoplight Results

Figure 4-1 presents the results from the Center High-Mount Stoplight (CHMSL) event during the on-road testing. Overall, the percent missed detections ranged from 10 percent to 40 percent. Visual-manual tasks generally had higher missed detection rates (20% to 40%) than auditory-vocal tasks (12% to 22%) and Just Drive. Just Drive had the lowest missed detection rate of all tasks at 9 percent. Higher missed detection rates for the visual-manual tasks may have been due to the necessity for the participants to conduct these tasks by looking down to the task or otherwise away from the road scene. When the participants looked down to execute the visual-manual tasks, only peripheral vision may have been available to detect the CHMSL event. So, as expected, it was difficult to see a CHMSL event during the visual-manual tasks.

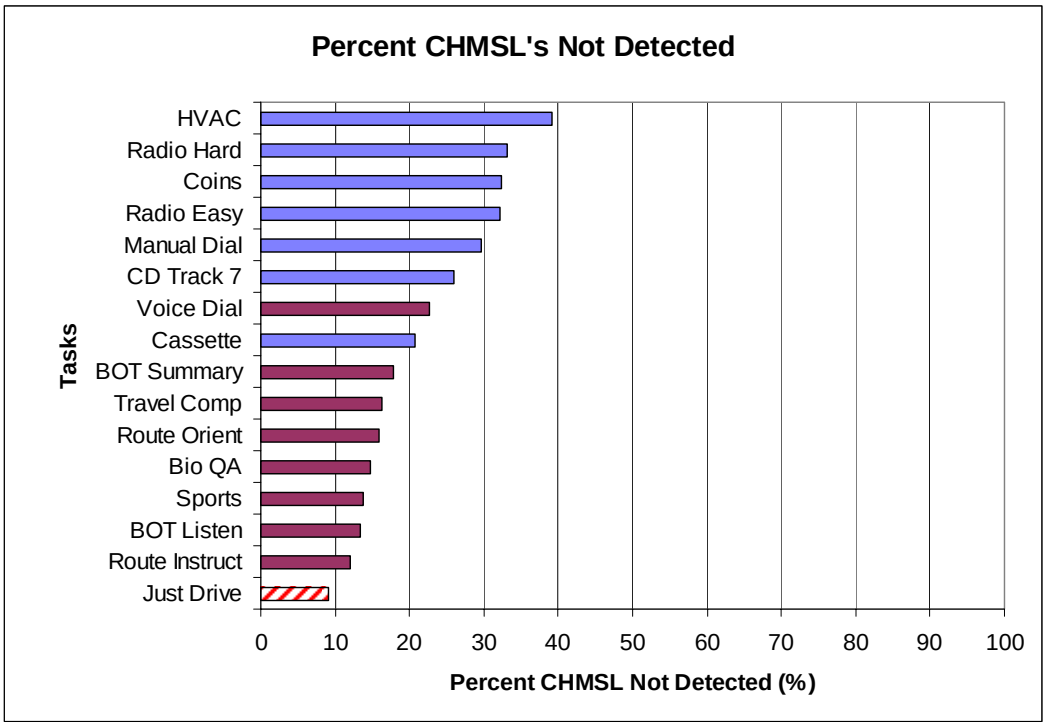


Figure 4-1. On-Road Percent of CHMSLs Not Detected by Task

For the CHMSL OED, the percentage of missed detections showed a good correlation with detection Response Time, $R^2 = 0.7221$ (see Figure 4-2). This was true for both the on-road and test track portions of the study.

Response time was calculated from the stimulus onset time to the participant's response. Responses less than 200 msec after stimulus onset were treated as anticipation responses and excluded. CHMSL and FVTS stimulus onsets began when a signal from the data acquisition system reached the follow or lead car, as appropriate.

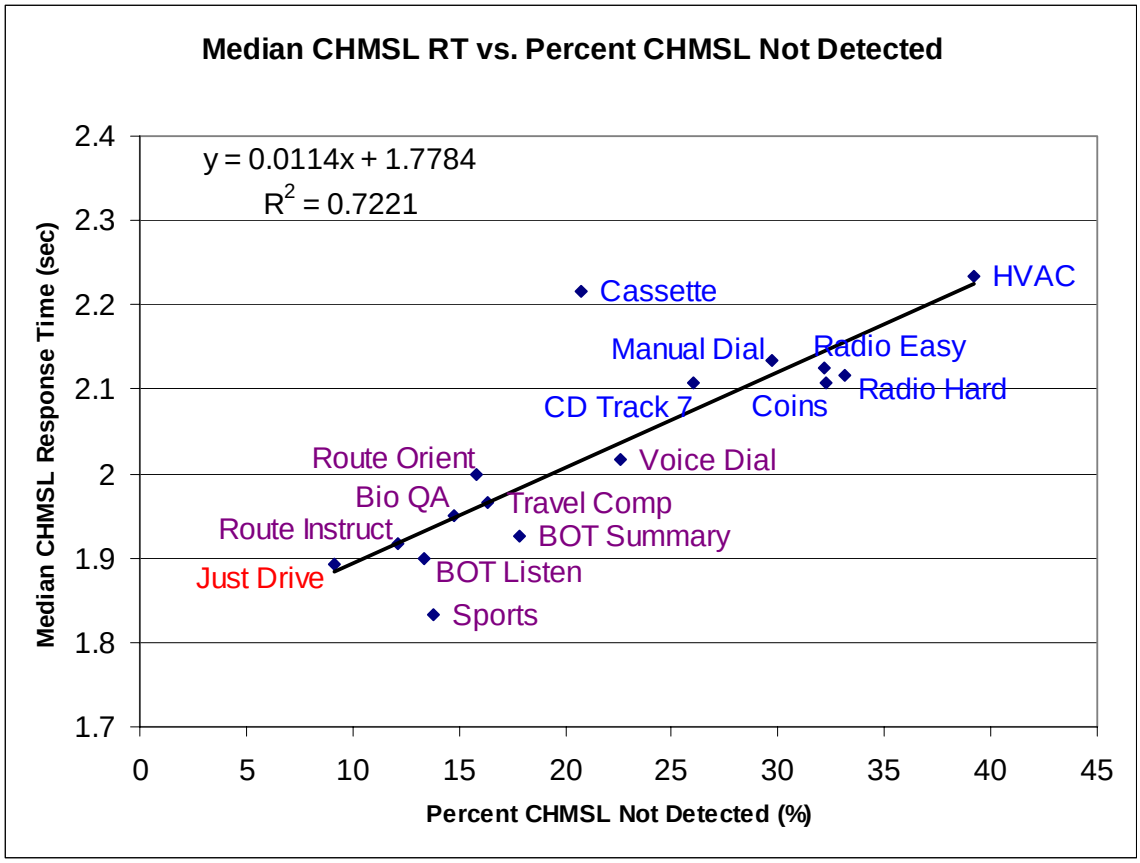


Figure 4-2. On-Road Percent Missed Detections Versus Response Time

For all of the OED scenarios, the general trends of the relationship percent missed detections and response times were:

- Tasks with **higher** OED event percentage missed were associated with **slower** response times
- Tasks with **lower** OED event percentage missed were associated with **quicker** response times

This result was as expected. Participants who detected most of the OED events, in general, responded more quickly, while participants who missed many OED events responded more slowly.

4.3.2 Lead Vehicle Deceleration Results

The results from the Lead Vehicle Deceleration (LVD) event during the on-road testing showed a similar pattern to that of the CHMSL event—the visual-manual tasks, in general, had a higher missed detection rate than the auditory-vocal tasks (see Figure 4-3), with the exception of the Insert Cassette task. LVD percent miss rates for visual-manual tasks were in the 25 percent to 45 percent range, while miss rates for the auditory-vocal tasks were in the 15 percent to 25 percent range.

For the LVD event, as was shown for the CHMSL event, the Just Drive task had the lowest or near-lowest percentage of missed detections. LVD stimulus onset was from receipt of a signal to the lead vehicle to disable cruise control and begin the coastdown, if driving conditions permitted it. A portion of LVD events were not detectable within the task length of short visual-manual tasks, but it was not possible to correct individual percent missed detections based on the detection threshold. See Chapter 8 for further discussion of this issue.

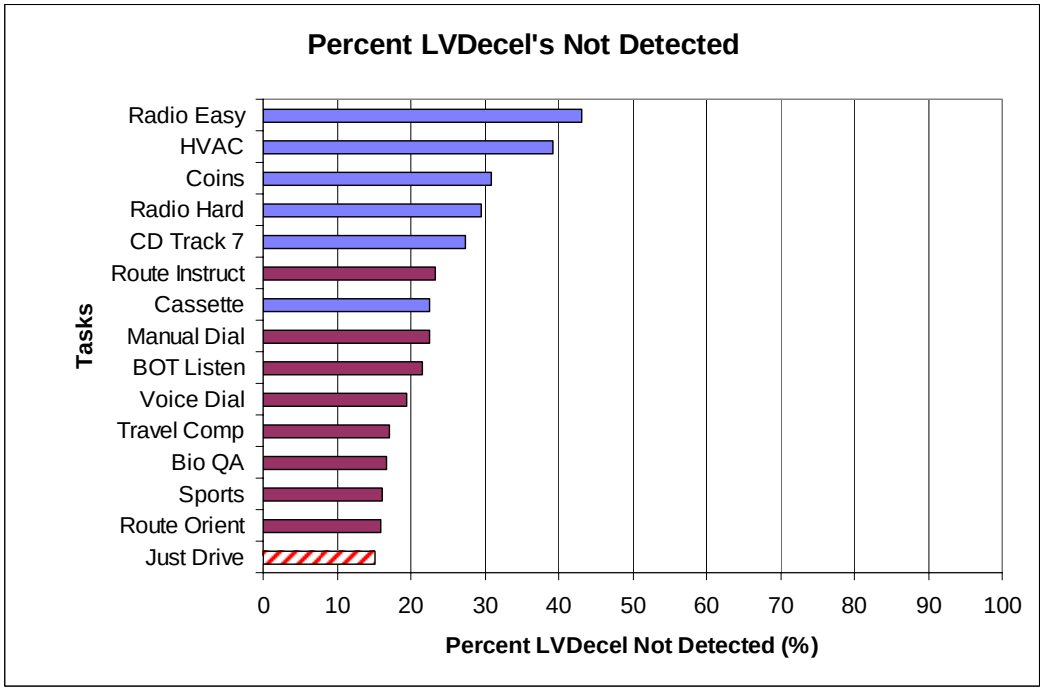


Figure 4-3. On-Road Percent Missed Detections for Lead Vehicle Decelerations

4.3.3 Follow Vehicle Turn Signal Results

The results from the Vehicle Turn Signal (FVTS) event during the on-road testing (Figure 4-4) showed a similar pattern to that of the CHMSL and LVD events: the visual-manual tasks had a higher missed detection rate than the auditory-vocal tasks. However, the visual-manual and the auditory-vocal tasks were interspersed. A visual examination of Figure 4-1, Figure 4-3, and Figure 4-4 suggests that interspersion for FVTS was greater than that for CHMSL and LVD events.

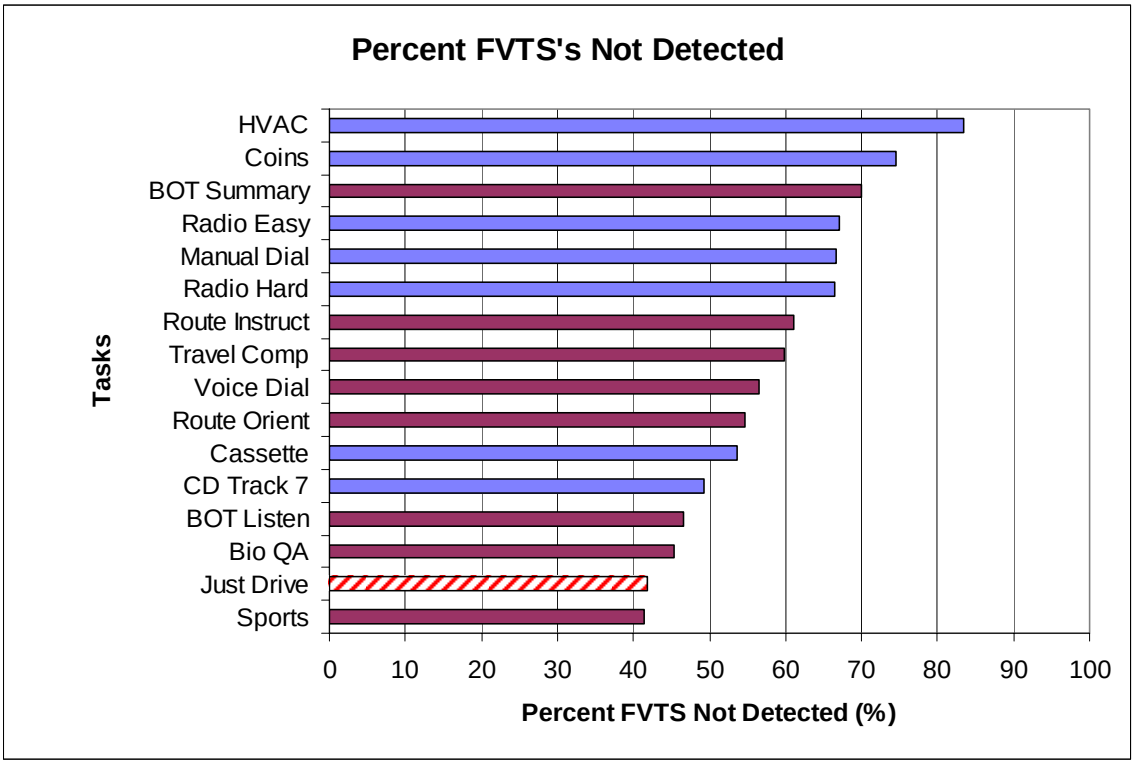


Figure 4-4. On-Road Percent Missed Detections for Follow Vehicle Turn Signal OEDs

Another discovery was that the missed detection rate for the FVTS event was much higher than both the CHMSL and LVD events—in the 40 percent to 90 percent range.

It was hypothesized that the FVTS event placed an unrealistic emphasis on events that occurred in the driver's rear view, and this would focus much more attention by the participant to the inside and outside rear view mirror than would occur in real-world driving. The much higher missed detection rate for the FVTS event than the other two OED scenarios shows that this hyper-focus on the rear view did not occur. Focus on the rear view mirrors was much less than the forward road scene when the participant was engaged in the experimental secondary tasks. During secondary task loading, drivers appeared to have prioritized the forward road scene over the scene to the rear.

4.3.4 Comparison of On-Road and Test Track Results

There was good correlation for percent missed detections between on-road and test track results for all three OED conditions: CHMSL, LVD, and FVTS.

For CHMSL, the correlation for percent missed detections between on-road and test track was $R^2 = 0.7961$; for LVD, $R^2 = 0.8881$; and for FVTS, $R^2 = 0.7211$ (see Figure 4-5, Figure 4-6, and Figure 4-7). These correlation values show a strong relationship between on-road and test-track results. Hence, there is good predictive value from OED track results to more realistic on-road conditions for these measures.

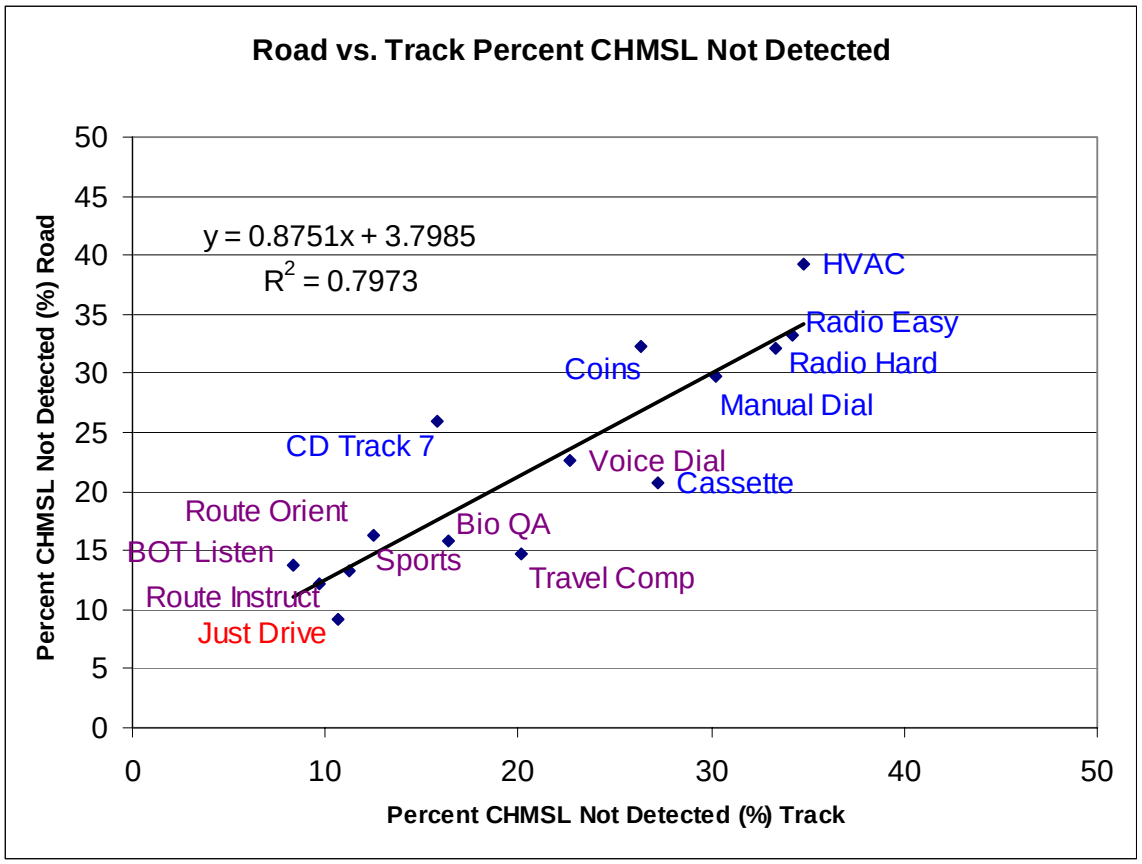


Figure 4-5. Comparison of On-Road Percent Missed Detections With Test Track

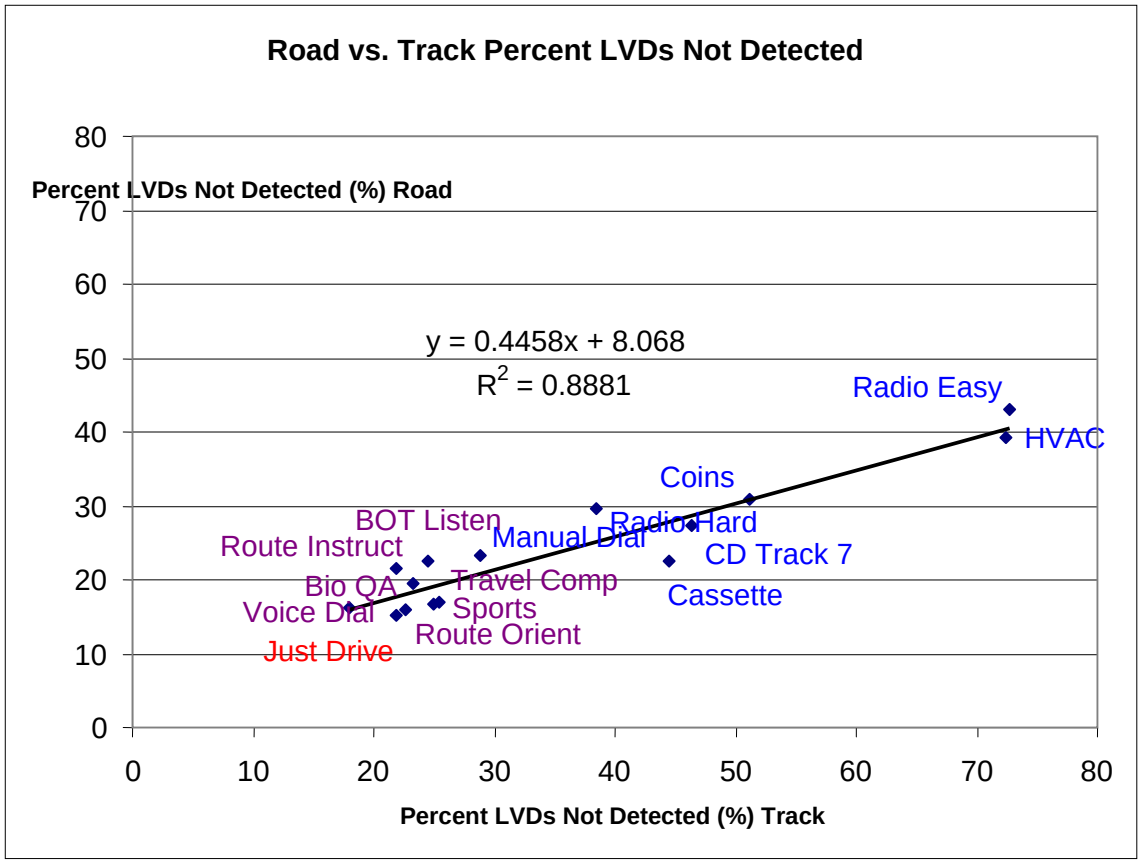


Figure 4-6. Comparison of On-Road Percent Missed LVDs With Test Track

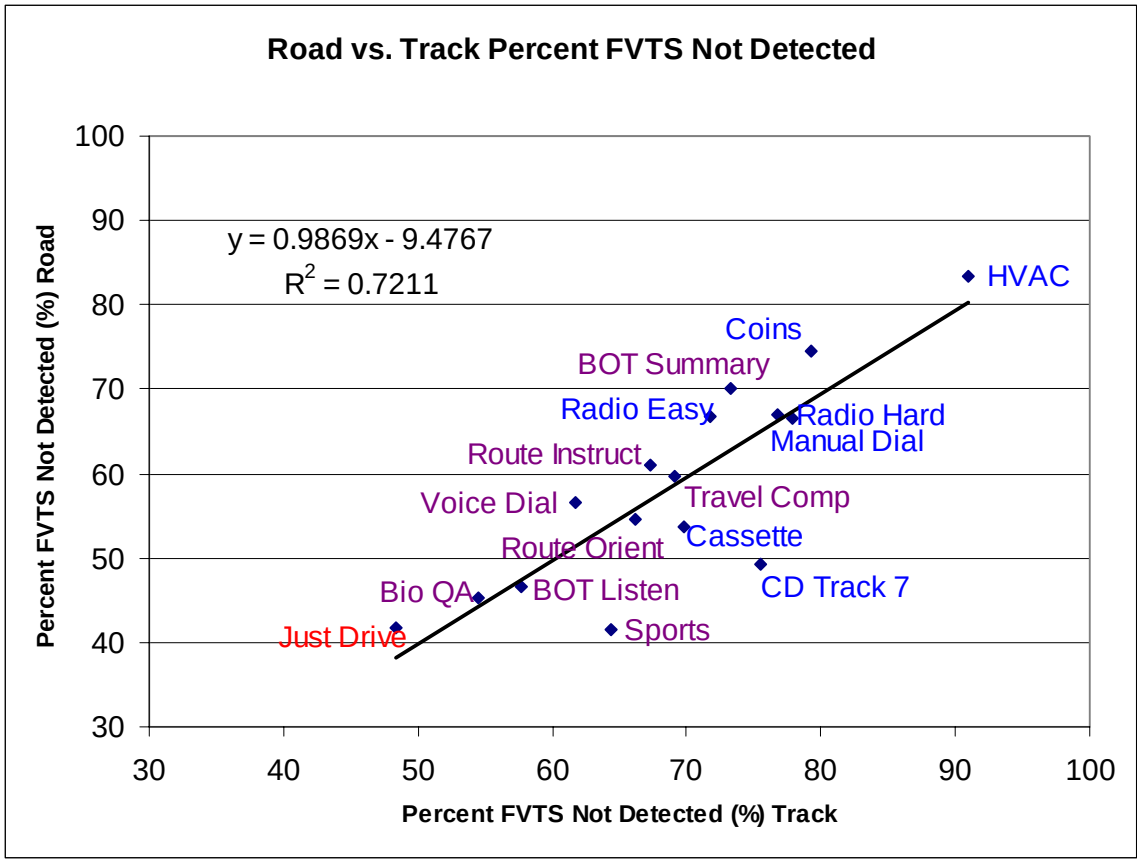


Figure 4-7. Comparison of On-Road Percent Missed FVTS With Test Track

4.3.5 Effect of Additional OED Response Window

The requirement for a test participant to respond within the boundaries of the task duration carried with it an implied workload. There was some concern that the OED response window, which ended when the task was completed, did not allow sufficient time for participants to respond. A test participant may actually have detected an OED stimulus within the task window, but may have been too busy with the task to respond within the time framework of the task. For such cases, responses could count as correct detections.

An investigation into OED response outside the boundary of task duration was conducted to find out if this was the case. The time epochs for responses were extended by 5 seconds, and OEDs beyond task end to this 5 seconds extension were tallied, and miss rates were re-calculated.

Table 4-2 shows the additional OED detections tallied in the 5 seconds extended response window. Visual-manual tasks, which may have caused the most missed detections due to the conflict of manual workload with OED response, are highlighted.

Table 4-2 also shows there were only a small percentage of OED detections tallied in the 5 seconds extended response window compared to the total number of trials: 1.53 percent and 1.22 percent for the CHMSL event in track and road conditions, respectively, and 0.78 percent and 1.27 percent for the FVTS event in track and road conditions, respectively.

Table 4-2. OED Detections in the Five-Seconds Extended Response Window

	OED detected in +5sec			
	Track		Road	
Task	**CHMSL**	**FVTS**	**CHMSL**	**FVTS**
Coins	4	6	8	3
Cassette	6	1	4	4
HVAC	3	1	4	7
Radio("Easy")	5	2	7	7
ManualDial		1	1	2
TravelComp			1	1
RouteOrient				
VoiceDial				1
BookOnTapeListen			1	
JustDrive		1		1
BiographicQ&A			1	
RouteInstruct			1	
Sports				2
Radio("Hard")	3		4	3
CD/Track7			1	1
RouteTracing	4			
Delta				
BookOnTapeSummary	1		1	3
DestEntry				
Read("Easy")	1	1		
Read("Hard")	1	3		
Map("Easy")				
Map("Hard")	3			
Total	31	16	34	35
Total trials	2024	2040	2776	2746
%	1.53	0.78	1.22	1.27
Probe ON beyond task	0	8	2	4

Figure 4-8 gives an illustrative example of the difference the missed detection rates with and without the five second extended response window. The difference in the two conditions was

Road CHMSL Miss Rates:
Task End Versus Task End+5 sec

illustratively small.

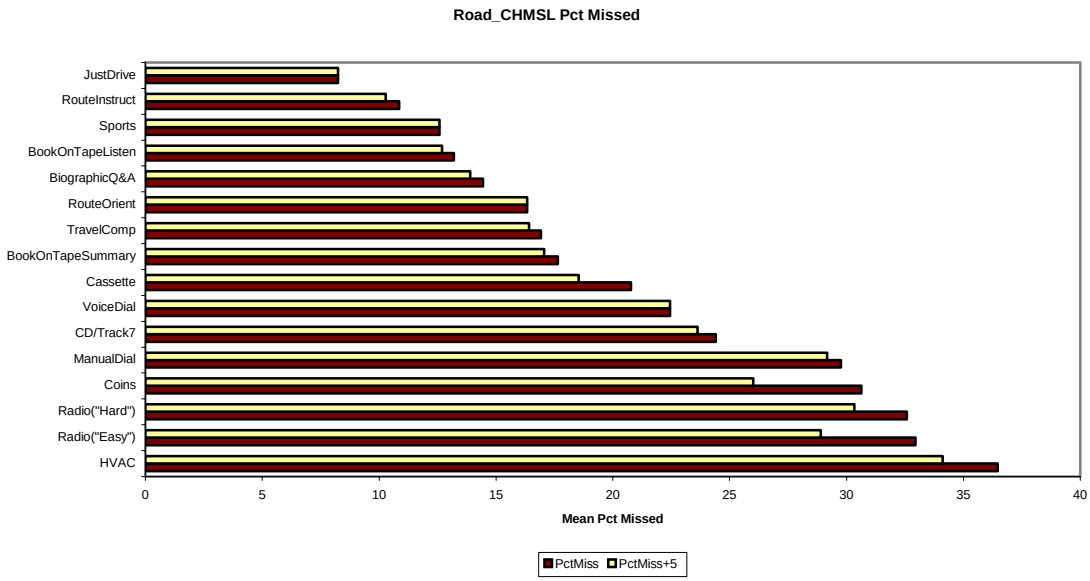


Figure 4-8. On-Road CHMSL Percent Missed Detections for End Task Versus End Task Plus Five Seconds

Figure 4-9 shows the same small differences in the missed detection rates with and without the 5 second extended response window for the FVTS OED event for the track results. These results of small differences in the missed detection rates with and without the 5 seconds extended response window were consistent for both CHMSL and FVTS OED events both on the road and on the test track.

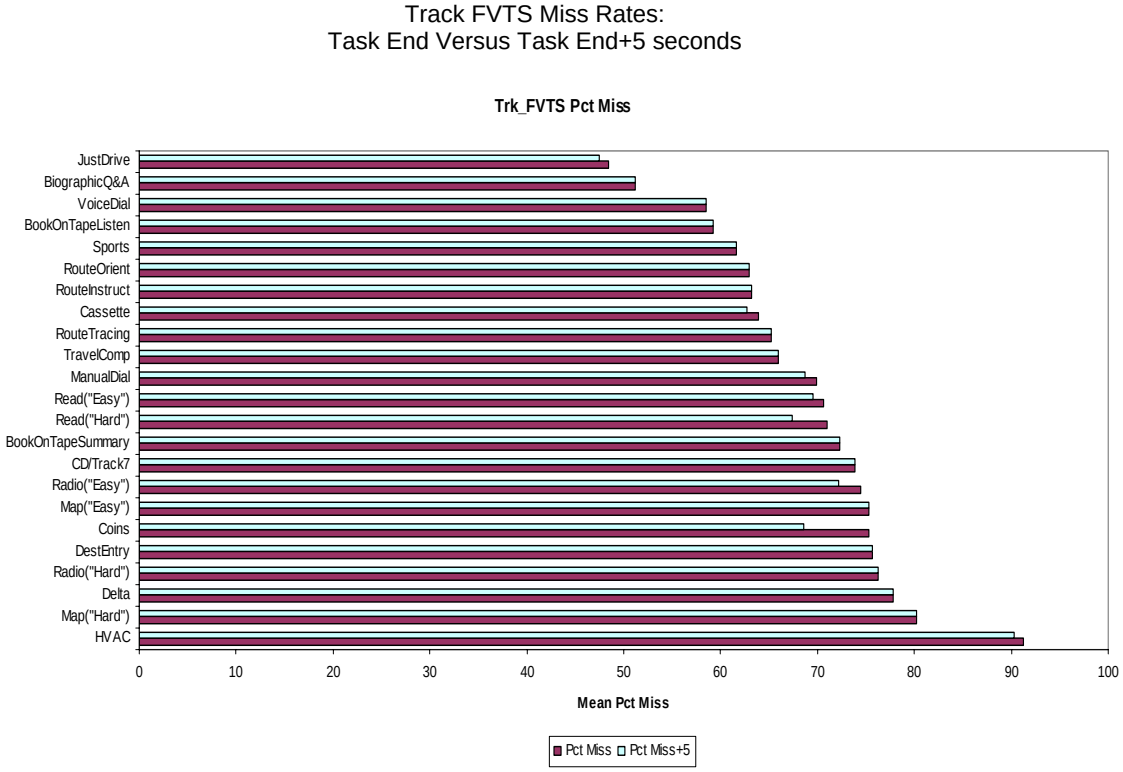


Figure 4-9. Track FVTS Percent Missed Detections for Task End Versus Task End Plus Five Seconds

4.3.6 Summary of OED Results

For all three OED scenarios, CHMSL, LVD, and FVTS, both the on-road and track testing showed that the visual-manual tasks had higher missed detection rates than the auditory-vocal tasks. This may have been due to the necessity for the participants to conduct these visual-manual tasks in a head-down manner, and look at the center part of the instrument panel or the center console to complete the tasks. Another contributing factor to the result showing higher missed detection rate for the visual-manual tasks than the auditory-vocal tasks was the inherently shorter duration of the visual-manual tasks. Only one OED presentation was scheduled during each shorter visual-manual task. However, a participant could receive up to three OED presentations (one of each type) during a longer auditory-vocal task, and the Destination Entry task. (However, on the test track, there was only one event presented during both auditory-vocal and visual-manual tasks, yet results correlated highly between venues in spite of differences in the number of events presented per task.)

For all of the OED scenarios, the percentage of missed detections showed a good correlation with Detection Response Time. The general trends of the relationship between percent missed detections and response times were that tasks with higher OED event percentage missed were associated with slower response times, and tasks with lower OED event percentage missed were associated with quicker response times.

For all of the OED scenarios, the Just Drive condition had the lowest or near-lowest percentage of missed detections.

The results from the FVTS event showed much greater interspersion between the visual-manual and the auditory-vocal tasks than in the other two OED scenarios. Another difference was that the missed detection rate for the FVTS event was much higher than both the CHMSL and LVD events. This showed that focus on the rear view was much less than the forward road scene when the participant was engaged in the in-vehicle tasks. During secondary task loading, drivers appeared to have prioritized the forward road scene over the rear visual scene.

There was good correlation for percent missed detections between on-road and track results for all three OED conditions. These correlation values showed a strong relationship between on-road and track results, which leads to a good predictive value from OED track results to more realistic on-road conditions.

There was some concern that the OED response window, which ended when the task was completed, may not have allowed sufficient time for the participant to respond, even if the OED had been detected. Therefore, an investigation into OED response outside of the boundary of task duration was conducted to find out if this was the case. The time epochs for responses were extended 5 seconds, OEDs beyond task end to this 5-second extension were tallied, and miss rates were re-calculated.

There were only a small percentage of OED detections tallied in the 5-second extended response window compared to the total number of trials: 1.53 percent and 1.22 percent for the CHMSL event in track and road conditions, respectively, and 0.78 percent and 1.27 percent for the FVTS event in track and road conditions, respectively (Table 4-2). These results of small differences in the missed detection rates with and without the 5-second extended response window was consistent for both CHMSL and FVTS OED events both on the road and on the test track.

4.4 Road Task Effects on Eyeglance Behavior

Driver eyeglance behavior was examined in the on-road venue using the same metrics as were examined in the test track venue (again through the efforts of Carol Flannagan at the University of Michigan Traffic Research Institute). As a result, an overview of the methods will not be repeated here (see “Test Track Task Effects on Glance Behavior” in Chapter (3). However, when comparing graphs in this section to those in the corresponding section in Chapter 3, it should be kept in mind that fewer tasks were administered in the on-road venue, so fewer tasks are plotted in the figures in this chapter.

4.4.1 Task Effects on Eyeglance Metrics

Table 4-3 shows the results of the linear mixed-model analyses performed across a variety of metrics using the on-road eyeglance data. As is apparent, there was a significant main effect of Task on all of these glance metrics, as well as significant main effects of Location Type, and a significant Task by Location interaction.

Table 4-3. Linear Mixed-Model Effects for Glance Metrics

The red triangles indicate effects significant at p < 0.05.

	← Metrics →								
Venue	**Effect**	**No. of Glances**	**Total Time At Location**	**Max Dur**	**Min Dur**	**Mean Dur**	**Med Dur**	**St Dev Du**	**Glance Rate**
Road	Gender								
	Task	▲	▲	▲	▲	▲	▲	▲	
	AgeGroup								
	Locat	▲	▲	▲	▲	▲	▲	▲	
	Task*AgeGroup								
	Task*Locat	▲	▲	▲	▲	▲	▲	▲	
	AgeGroup*Locat								

For ease in comparing the Task by Location Type interaction effects from the road data for these (and some other) variables to the same findings from the test track (in Chapter 3), the graphs and corresponding explanations are presented in the following subsections.

4.4.1.1 Number of Glances

Figure 4-10 shows the interaction of Task by Location Type for the on-road venue on the metric of Number of Glances. As was true for the data from the test track, there was a great deal of variation across tasks in Number of Glances per task. This is consistent with expectations based on variation in task durations (since longer task durations would allow more glances to be made). However, the variation across tasks also was a function of the type of location at which gaze was directed (road, situation awareness, or task-related), as was true for the test track.

Within the interaction of Task by Location Type, there were several sub-patterns of interest, similar to those in the test track data. Focusing first on glances to the road, the region to which the highest number of glances were made, the Just Drive task led to more than 30 glances, on average (as was true for the test track). Also, as was true for the test track, a number of the auditory-vocal tasks produced nearly as many glances as Just Drive, between 20 and 35. Auditory-vocal tasks appear toward the right side of the figure and are highlighted with a dark red bar beneath the task names and numbers. One auditory-vocal task, Book-on-Tape Summarize, had fewer glances than all other auditory-vocal tasks (averaging about 35 seconds rather than 2 minutes). For the Just Drive and auditory-vocal tasks, the number of glances to the situation awareness category paralleled the number to the road and the points were virtually on top of the points for the number of glances to the road. Thus, the patterns of glancing to the road and situation awareness location types for auditory-vocal tasks resembled the pattern for Just Drive, just as it did for the test track data.

The pattern for visual-manual tasks was again distinct, just as it was on the test track. First, the number of glances to the road was dramatically lower for most of the visual-manual tasks (there were one-sixth to a little over half of the number of glances to the road for Just Drive). This is to be expected based on task duration alone. Shorter tasks would allow fewer glances to be made overall. In addition, for visual-manual tasks, glances to the situation awareness category were fewer than to the road, less than half the number to the road on average. The number of task-related glances was slightly less than those to the road, but tended to closely track the road glances. Thus, for visual-manual tasks, on the road (as on the track), drivers usually glanced back and forth between the road and the task. Of course, there was occasionally a glance at the mirrors.

In Chapter 3, it was pointed out that auditory-vocal tasks are not normally thought of as requiring any task-related glances, yet there were in fact a few glances scored as task-related. These tended

to be glances up, either to the rear-view mirror in some of the language-production tasks (such as Biographical Q&A), or glances up to the visor area. As mentioned in the discussion of the eye data from the test track, these were very carefully scored and, in particular, the glances to the mirrors associated with auditory-vocal tasks were carefully discriminated on the basis of different body movement, head movement, and eye movement cues from glances to the mirror for checking traffic. Only those glances that were discriminable from typical mirror-checking to monitor traffic were scored as related to auditory-vocal tasks. As mentioned previously, one hypothesis is that when a task requires the use of working memory, drivers look up, as if to visualize or "look at" the contents of working memory—or with language-production tasks, may seek a listener to look at out of habit. Indeed, in the on-road data, a few, infrequent upward glances were observed, just as they were in the test track data, as shown by the yellow line in the area of the auditory-vocal tasks. These glances were associated with those auditory-vocal tasks for which retrievals from long-term memory, mental calculation, rehearsal in memory, or generation of linguistic material were needed.

Note: Further details on the scoring of upward glances are provided in the corresponding eyeglance section of Chapter 3.

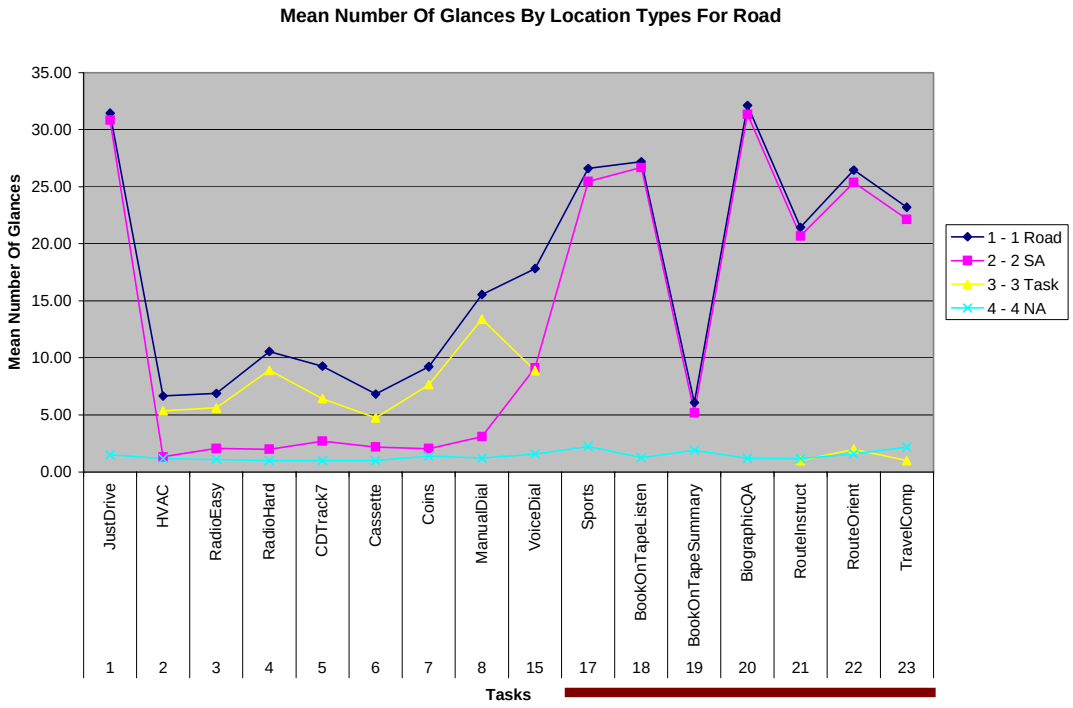


Figure 4-10. On-Road Mean Number of Glances by Task and Location Type

4.4.1.2 Glance Duration (by Task and Location Type)

Figure 4-11 shows Median Glance Durations, Figure 4-12 shows Mean Glance Durations, and Figure 4-13 shows Maximum Glance Durations. All three depict similar patterns. As was true for the test track data, the durations of glances to the road were much longer than glances to other regions, but particularly for the Just Drive and auditory-vocal tasks. For the Just Drive task, glances to the road tended to be somewhat shorter than for the test track, clustering in the range represented approximately by 2.6 seconds (median) to 3.6 seconds (mean) in duration and contrasting with the test track data of 6 seconds (median) to 8 seconds (mean) in duration. For auditory-vocal tasks, in which the eyes could be forward and on the road for the entire task, glances to the road tended to be even longer. For the road data, most of them fell in the range from about 3 to 6 seconds based on medians, or from 4 to 8 seconds based on means. This contrasted with the test track ranges from about 7 to 14 seconds, based on medians and from about 9 to 16 seconds, based on means with the mixed-mode Voice Dial task falling in between the range for visual-manual tasks and Just Drive. For all visual-manual tasks, glance durations to the road and all other areas, tended to be under 2 seconds in duration. Figure 4-14 shows mean glance durations for only task-related and situation-awareness glances, so that the scale of the figure could be enlarged within the region of these glance durations. With this scale change, it is possible to see that most task-related glances are between 0.80 seconds and 1.20 seconds for the road, which is similar to that for the test track range of 0.80 seconds and 1.40 seconds, while most situation-awareness glances averaged between 0.4 and 0.7 seconds in duration as was also true for the test track.

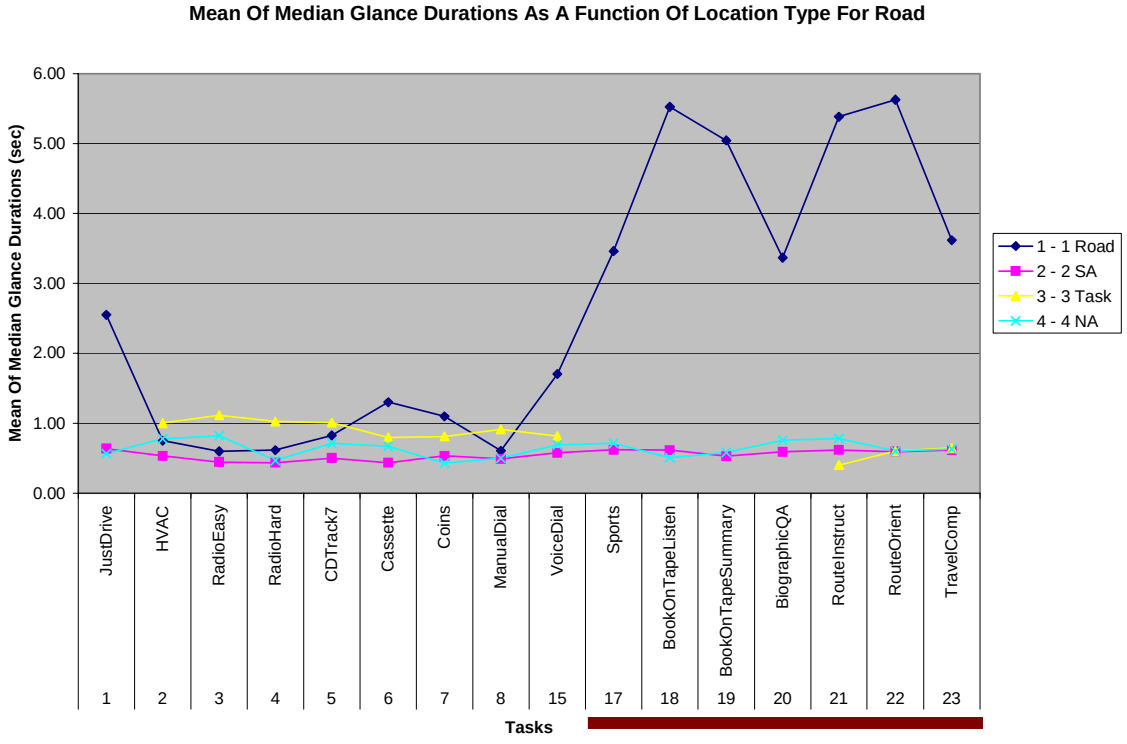


Figure 4-11. Road Mean of Median Glance Durations by Task and Location Type

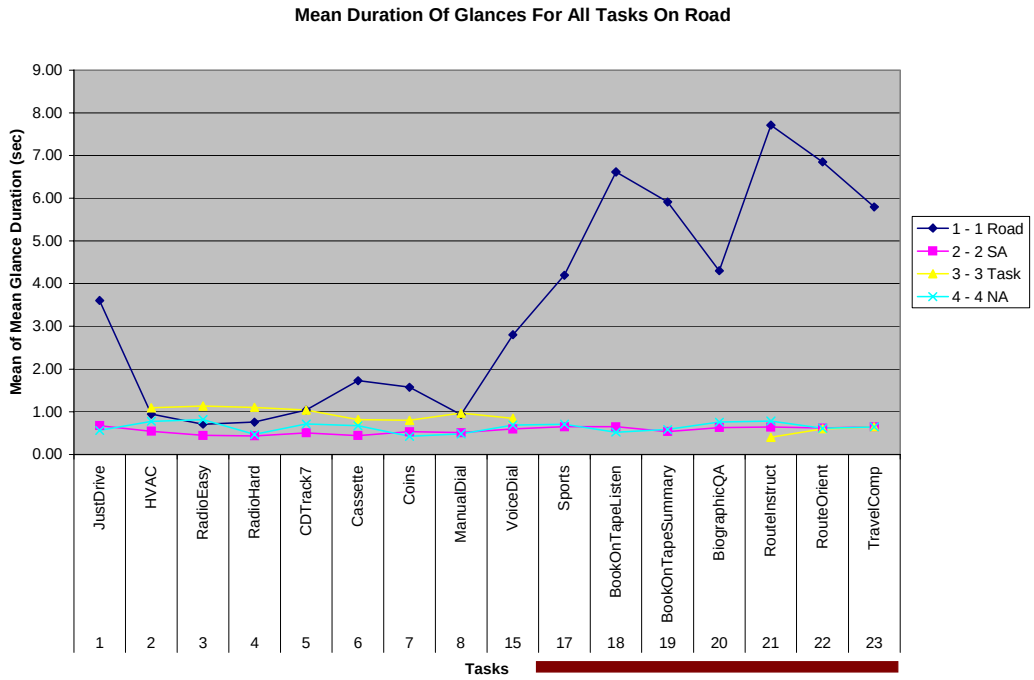


Figure 4-12. Road Mean of Mean Glance Durations by Task and Location Type

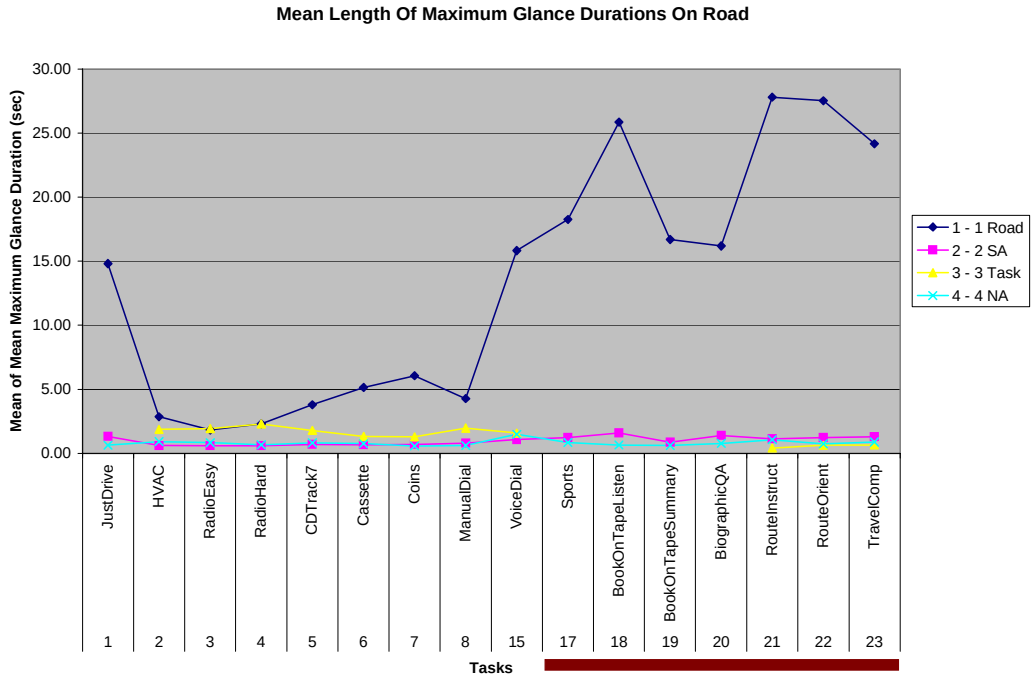


Figure 4-13. Road Mean of Mean Maximum Glance Durations by Task and Location Type

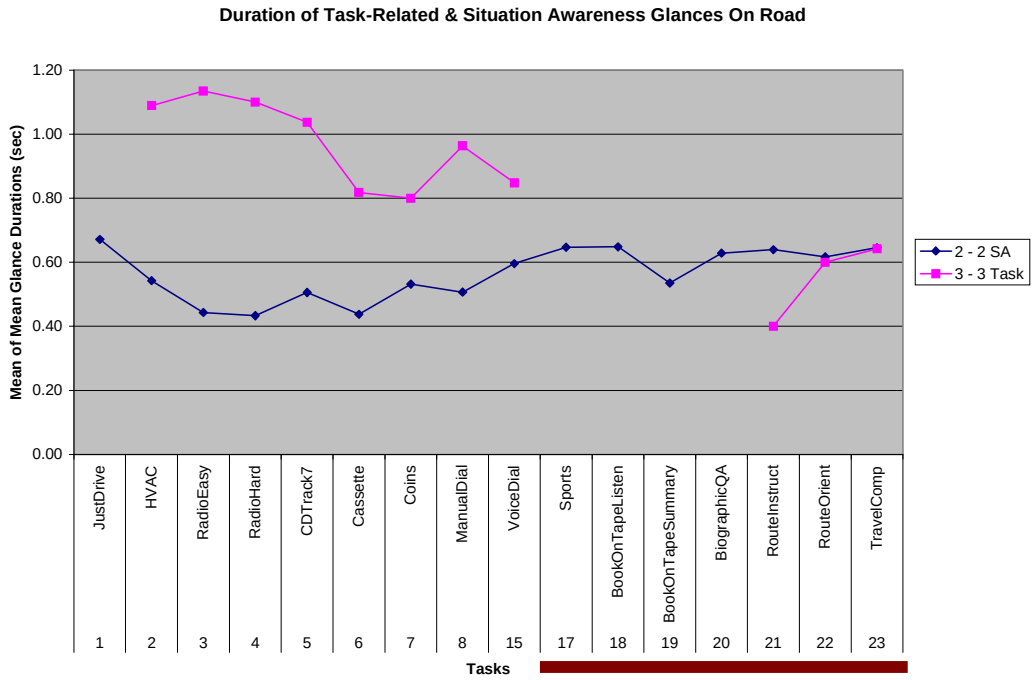


Figure 4-14. Road Mean of Mean Duration of Task-Related and Situation Awareness Glances

Figure 4-15 shows the interaction effect of Task by Location on Minimum Glance Duration. This figure shows that the shortest task-related glances on the road for the visual-manual tasks were just under 0.50 seconds in duration, longer than the minimum glances to the road and situation-awareness areas. The exceptions were the minimum task-related glances upward associated with auditory-vocal tasks, which tended to range from 0.40 to 0.63 seconds in duration. Also, the minimum glances to the road for Book-on-Tape-Summarize and Route Instructions tended to be a bit longer than for other tasks.

4.4.1.3 Glance Rate (for Task by Location Type Interaction)

Figure 4-16 shows the interaction of Task by Location Type on the Rate of Glancing to Each Location Type. As was true for the test track data, visual-manual tasks produced the highest glance rates to the road (in the range from 0.4 to 0.6 glances per second), as contrasted with the range from just under 0.2 to 0.3 per seconds for Just Drive and the auditory-vocal tasks. The glance rate to mirrors and speedometer (situation awareness) fell within the range from 0.10 to 0.30 glances per second, with the visual-manual tasks falling lower in the range. The glance rate to task areas for visual-manual tasks varied between 0.30 and 0.50, dropping off dramatically for the mixed-mode Voice Dial task.

Glance rates must be interpreted with caution, as indicated in Chapter 3. Simply because Task A had a higher glance rate to the Road (e.g., 0.5 glances per second) than Task B (e.g., 0.25 glances per second) does not necessarily mean that Task A was associated with higher levels of Road monitoring than Task B. This is especially true for tasks with fewer but longer glances to the road. Glance rates must be considered in conjunction with the number of glances and duration of glances made during a task.

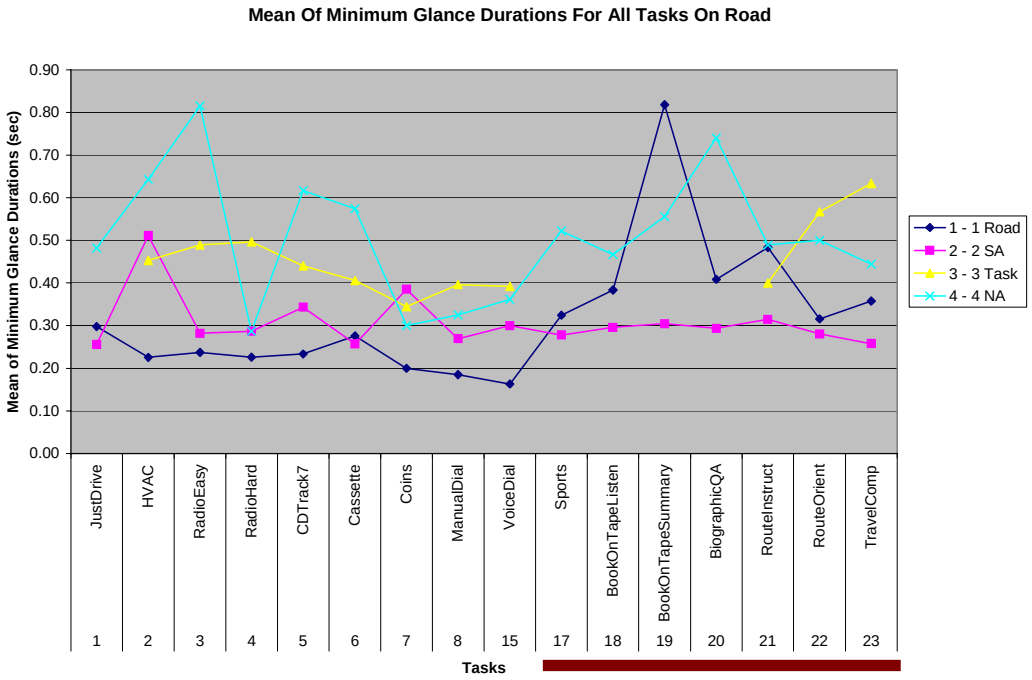


Figure 4-15. Effect of Location Type on Mean of Minimum Glance Durations

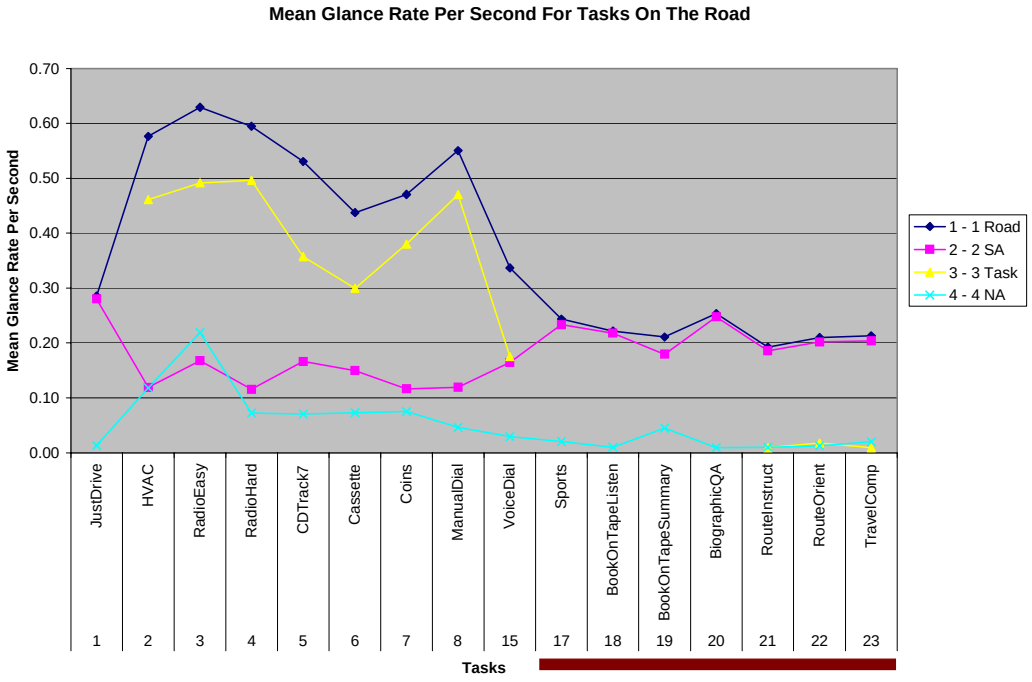


Figure 4-16. Road Rate of Glancing to Each Location Type by Task

4.4.1.4 Proportion of Task Duration Spent Looking at Each Location

Figure 4-17 shows the Task by Location Type interaction for the road data on the metric of Proportion of Task Duration Spent Looking in Each Location Type by Task.

As mentioned in Chapter 3, proportions are most comparable when they relate to the same or similar durations. DWM task durations can differ considerably from task to task, especially among visual-manual tasks. Proportions mask duration differences even though those differences may be important. Proportions can be misleading when the task durations are substantially different. Therefore, caution is urged in the interpretation of this type of measure.

The Proportion of Task Duration Spent Looking at the Road Location (shown in blue on the graph) again discriminated very well between tasks that were visual-manual and those that were primarily auditory-vocal, just as with the test track data. The mixed-mode Voice Dial task, which had some visual-manual elements and some auditory-vocal elements, fell in between the visual-manual and auditory-vocal tasks, with Voice Dial resembling the visual-manual pattern more closely, which was also true in the test track data. As with the test track data, the auditory-vocal tasks also looked similar on this measure to the Just Drive task, though even more time was spent looking at the road when performing them than when just driving (0.87 versus 0.81). During visual-manual tasks, the proportion of time during a task that drivers looked at the road averaged only 0.53.

The Proportions of Task Duration Spent Looking at Situation Awareness Locations (mirrors and speedometer), varied over a narrower range across all tasks, with the Just-Drive task showing a slightly higher proportion of time on mirrors than when drivers were engaging in a an additional in-vehicle task (Figure 4-17). In Figure 4-18, data for just the mirrors is shown in a similar manner (with glances at the speedometer removed for this analysis), using the same measure. However, the scale has been enlarged, so that the magnitude of the effects can be compared. Relative to the Just-Drive task, visual-manual tasks led to a larger drop in mirror-viewing, on average, than did auditory-vocal tasks. On average, for Just Drive, mirrors were viewed for 15.4 percent of a task's duration on the road (versus 14.3% on the test track). On average, for auditory-vocal tasks, it dropped slightly to 11.4 percent on the road (versus 10.7% on the test track) and for visual-manual tasks, it dropped further to 6.4 percent on the road (versus 7.96% on the test track). In the graph, these percentages are plotted as their corresponding proportions: 0.154, 0.114, and 0.064.

As was done on the test track data, for the road data, additional analyses were also conducted using linear mixed-model analyses to examine whether breadth of scanning narrowed under higher-workload tasks (regardless of type). The outcome of these analyses confirmed that for the road data, as for the test track data, there were significant differences (at $p < 0.05$) in mirror scanning behavior between tasks classified as high- and low-workload for multiple measures examined on glances to the mirror location. In fact, eight out of eight measures examined for mirror glances on the road showed significant effects (versus only four of eight for the test track). These measures were: number of glances to the mirrors, total glance time to the mirror location, percent (proportion) of time during task spent viewing the mirror location, and maximum glance duration to the mirror location—all of which also showed significant effects on the test track. In addition, however, in the road data, the measures of Mean Glance Duration, Median Glance Duration, Minimum Glance Duration, and Standard Deviation of Glance Durations for glances to the mirror locations also showed significant effects.

The Proportion of Task Duration Spent Looking at Task-related Areas (shown in yellow in Figure 4-17) depended heavily on the nature of the task, and was primarily related to the visual-manual tasks, ranging from about 0.24 to 0.52. This measure for the visual-manual tasks provided an overall indication of how much of the task period was spent looking at the task (and, as discussed under the test track findings, may perhaps be an overall indicator of visual demand). Rank order of visual demand for the subset of tasks tested on the road (based on proportion of time spent looking at task-related areas) were: (1) Radio (Hard and Easy), (2) HVAC, (3) Manual Dial, (4) CD/Track 7, (5) Coins, (6) Insert Cassette, and (7) Voice Dial. Auditory-vocal tasks, though generating some task-related glances up to the headliner/visor and/or rearview mirror areas, approached proportions of 0.00.

There were fewer visual-manual tasks for which the proportion of time spent viewing the task exceeded time spent viewing the road (or was equal to it), and the correspondence with higher miss rates for CHMSLs was not as straightforward as it appeared to be in the test track data. These tasks can be identified in the graph where the task-related (yellow) line is above the line for road glances. As mentioned previously, the relationship between proportion of task time spent viewing task versus road, and its correspondence with the staging of an event may be worth exploring further in future research on improved measures of visual demand.

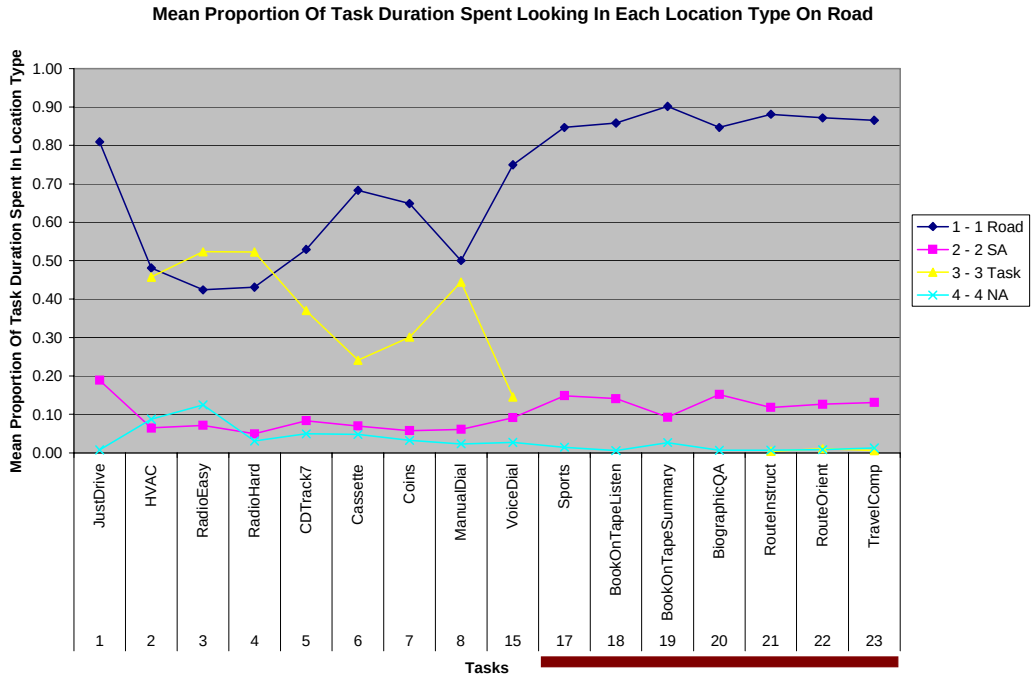


Figure 4-17. Road Mean Proportion of Task Duration Spent Looking at Each Location Type by Task

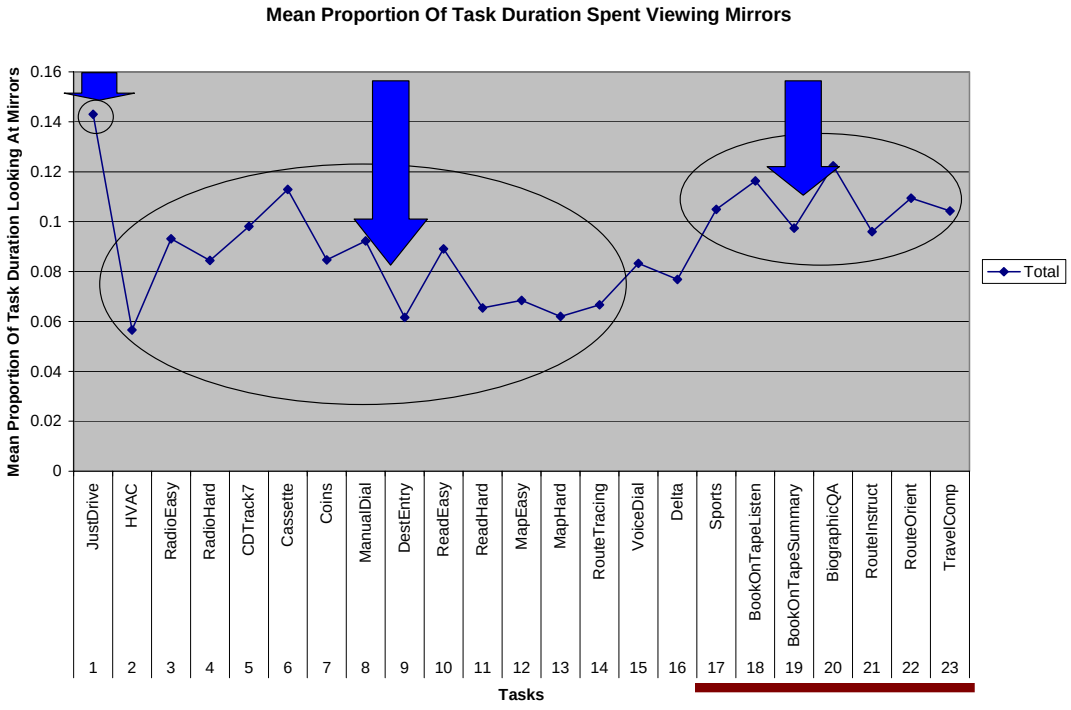


Figure 4-18. Road Mean Proportion of Task Duration Spent Viewing Mirrors

Note 1: Compare with similar line for situation awareness locations (mirrors and speedometer) shown in pink in Figure 4-17.

Note 2: Scale enlargement in this figure permits the size of effect to be compared between Just Drive (first circle), visual-manual tasks (second circle, longest downward arrow), and auditory-vocal tasks (third circle, shorter arrow). Visual-manual tasks led to larger drops in mirror viewing than did auditory-vocal tasks (relative to Just Drive).

If the long glance durations and periods of glancing at the road were indicative that drivers were inattentive during auditory-vocal tasks, then measures of attentiveness to event detection should indicate that higher percentages of events were missed during these auditory-vocal tasks. Specifically, Figure 4-19 should show an increase in percent CHMSLs missed on the road.

Note: It is important to remember that Figure 4-19 and others like it are based on data from the 18 research participants from whom eye data were reduced for the road venue. They, therefore, represent only a sub-sample of the larger data set on event detection described earlier in this chapter.

The points in Figure 4-19 that were associated with auditory-vocal tasks showed a slight increase in percent missed CHMSLS over Just Drive. Averaging over auditory-vocal tasks, 6 percent of CHMSLs were missed (represented by blue upward-pointing arrow in Figure 4-19), versus 0 percent for Just Drive. This was the same trend as seen on the test track, though these CHMSL miss rates were lower than for the corresponding conditions on the test track (10.4% missed CHMSLs during auditory-vocal tasks on the test track versus 6% for Just Drive on the test track). Furthermore, there was an even greater increase in percent missed CHMSLs for visual-manual tasks than for auditory-vocal tasks (23% missed CHMSLs on the road represented by the taller blue arrow), the same trend as seen on the test track (though on the test track 19.5 percent of

CHMSLs were missed). Thus, while slightly more CHMSLs were missed during auditory-vocal tasks than during Just Drive, many more were missed during visual-manual tasks (both on road and test track).

Similarly, as discussed in the test track data, if drivers were inattentive during auditory-vocal tasks, Figure 4-20 should show slower Response Times associated with CHMSL events that were detected during auditory-vocal tasks on the track. However, the pattern for Response Times to CHMSL events in Figure 4-20 showed very little difference between task types, on average. Response times on the road to CHMSLs for Just Drive were 2.09 seconds (versus 2.04 seconds for the test track), 2.18 seconds for auditory-vocal tasks (versus 2.09 seconds on the test track), and 2.18 seconds for visual-manual tasks (versus 2.05 seconds on the test track).

Together, these results from the road data for Percent Missed CHMSLs and Response Times to CHMSLs suggest that the concentration of gaze on the forward roadway observed in drivers performing auditory-vocal tasks on the road (as with those observed on the test track) was associated with only very subtle changes in attentiveness to CHMSL events, an increase in miss rate from 0 percent for Just Drive to 6 percent for auditory vocal tasks, with little change in response times, and these changes were much less pronounced than those produced by visual-manual tasks.

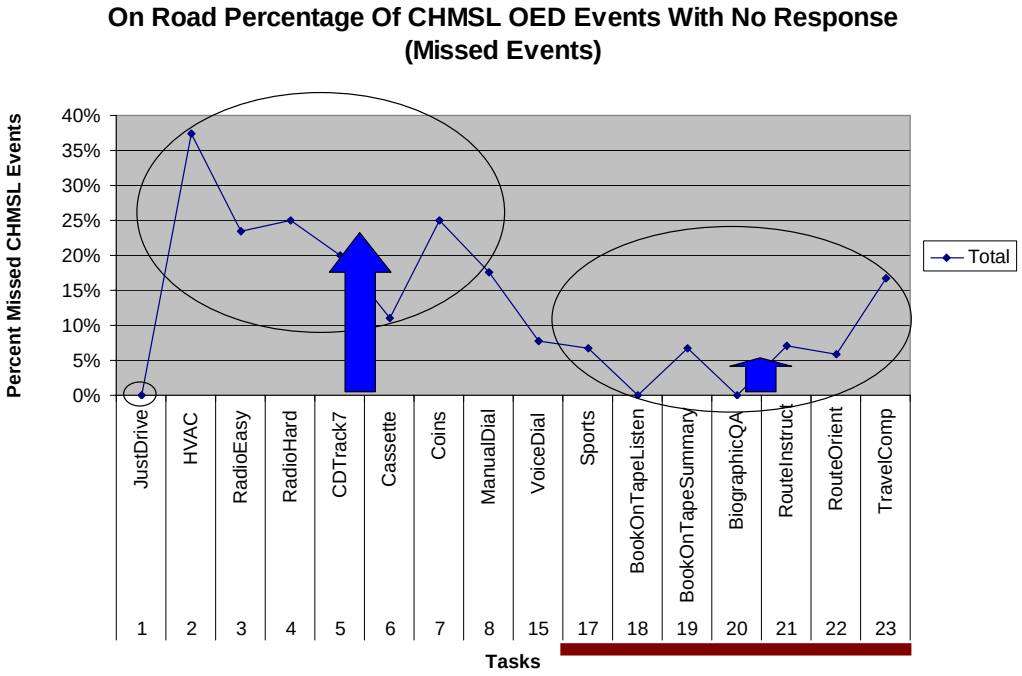


Figure 4-19. Percent CHMSLs Missed on the Road as a Function of Tasks (for comparison to glance patterns)

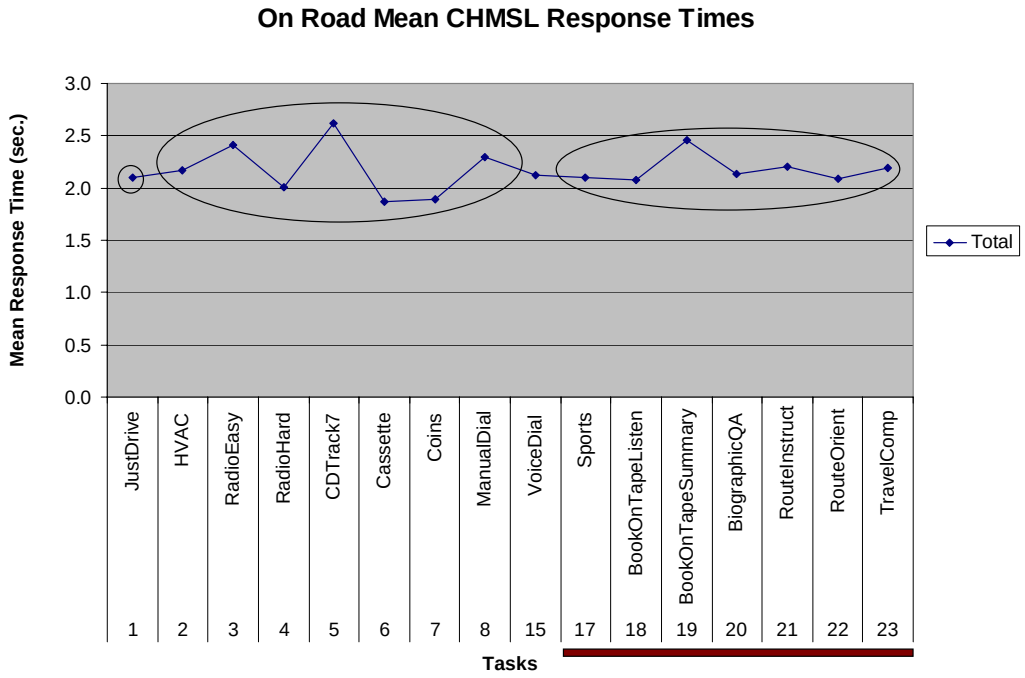


Figure 4-20. Response Times to CHMSLs on the Road as a Function of Tasks (for comparison to glance patterns)

The data on responsiveness to FVTS events is shown for the road data in Figure 4-21 and Figure 4-22 (related to FVTS). These data indicate that there was somewhat more intrusion from auditory-vocal tasks on detection of these peripheral events (which appeared in the left outside mirror). However, it was again less than that produced by the visual-manual tasks. Averaging across the auditory-vocal tasks shown in Figure 4-21, 29 percent of FVTS events were missed on the road during Just Drive (compared with 22% on the test track), versus 43 percent on the road for auditory-vocal tasks (compared with 45.43 percent on the test track). (The mean Percent FVTS mss rate is represented by the blue upward arrow shown above dark red bar at far right of figure.) This compared with 65 percent missed FVTS events on the road for visual-manual tasks, on average (indicated by taller blue upward arrow), compared to 63.54 percent missed FVTS events on the test track for visual-manual tasks. While the level of inattentiveness was still considerably less for auditory-vocal tasks than for visual-manual tasks, it was nonetheless more distinct for FVTS events than for CHMSLs, occurring mostly for the three most difficult auditory-vocal tasks. It was also consistent with the findings in Figure 4-18, indicating some reduced scanning of mirrors during auditory-vocal tasks (and even more reduced scanning for visual-manual tasks).

Response time data for the road (in Figure 4-22) indicated that during Just Drive, participants responded to FVTS events within 2.52 seconds, on average (versus 2.55 seconds for the test track). For auditory-vocal tasks, response times averaged 2.46 seconds for the road (versus 2.53 for the test track). Response times for visual-manual tasks averaged 2.60 sec; for the track the average response times were 2.088 seconds, but there was variability across the set of visual-manual tasks on the test track for response times to detected FVTS events).

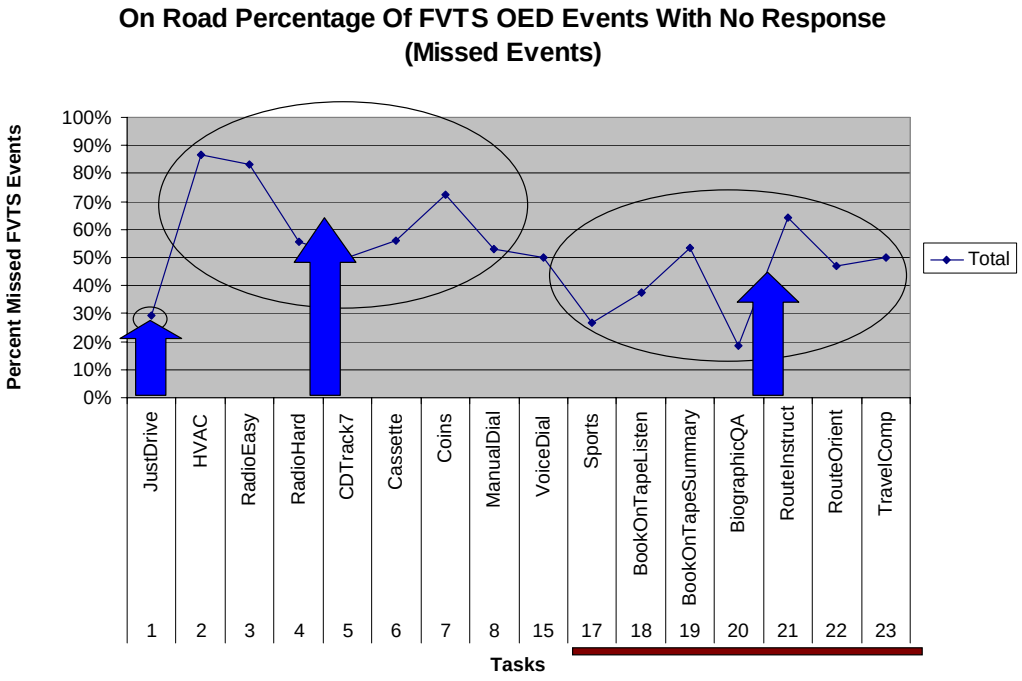


Figure 4-21. Percent FVTS Missed on the Road as a Function of Tasks (for comparison to glance patterns)

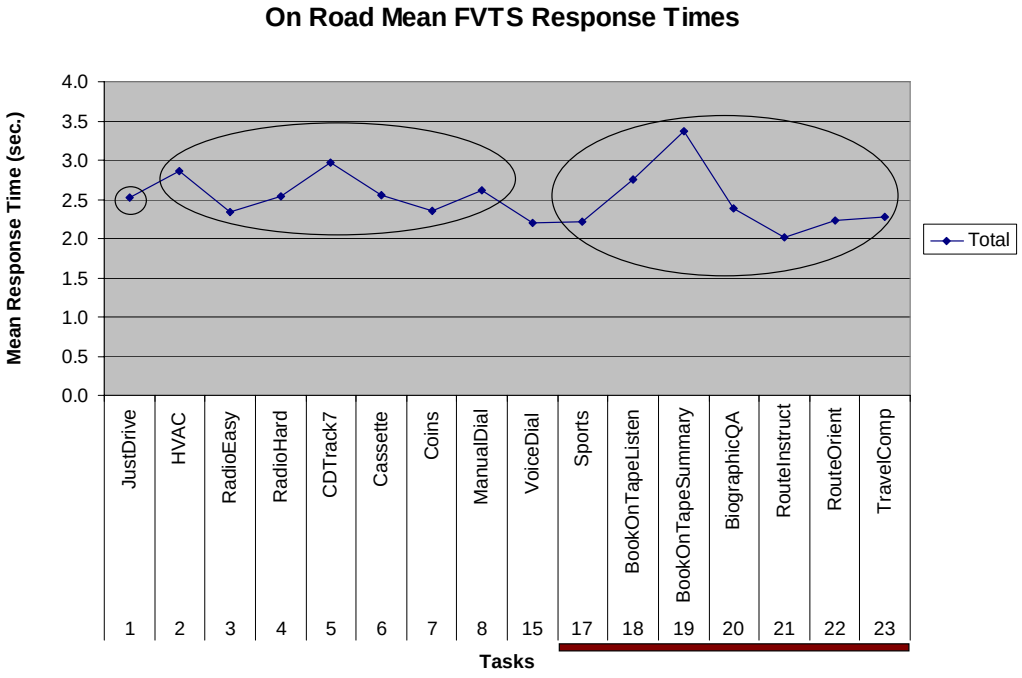


Figure 4-22. Response Times to FVTS Responded to on the Road as a Function of Tasks (for comparison to glance patterns)

The results for data from the road on responsiveness to LVD events are shown in Figure 4-23 and Figure 4-24. Both graphs present results similar to those presented for the CHMSL event.

> Note: Because the sample of data is small (18 participants), when it was decomposed by task, by glance location, and by whether or not an event was detected, there were some instances in which either no glances occurred, or no events could be launched (in the instance of short tasks, such as Book-on-Tape Summarize). When this occurred, it will be evident in the figures as "missing points" on the graph.

Auditory-vocal tasks (which were associated with a concentration of gaze on the forward roadway), were associated with a slight elevation in Percent Missed Lead Vehicle Decelerations on the road when compared to Just Drive (16% Missed LVDs for auditory-vocal tasks, on average, for the road versus 16.6 percent for the test track), compared with 12 percent for Just Drive on the road versus 13% on the test track). However, this slight elevation was again less than that seen for visual-manual tasks (28% Missed LVDs for visual-manual tasks, on average, on the road versus 33% on the test track), (though some of these had rates of missed LVDs for the methodological reason that they were too short for the event to even be detectable within the task's length). (Blue arrows depict these average miss rates in the figure). Response times to detected LVD events were 4.26 seconds, on average, for Just Drive done on the road (versus 5.25 seconds for the Test Track), 5.26 seconds for auditory-vocal tasks on the road (versus 5.45 seconds on the test track), and 5.41 seconds for visual-manual tasks on the road (versus 5.90 seconds on the test track). These patterns were consistent with the percent miss rates. Interestingly, response times on the road were faster to LVD events than they were on the test track.

Therefore, as with the test track, considering all of these results together, a hypothesis that very long glances to the forward roadway during the auditory-vocal tasks in this study may have indicated some level of inattentiveness, received little support from the data. The results from the road replicated those from the test track. As discussed previously, the magnitude of the effects was much smaller than might have been expected. Attentiveness to events was higher during auditory-vocal tasks than during visual-manual tasks. Effects for peripheral FVTS events were consistent with the notion that scanning of the periphery was shed to some degree during auditory-vocal tasks, as the eyes concentrated on the forward roadway more during auditory-vocal tasks. However, for visual-manual tasks, it appeared that scanning of the periphery was shed to a much greater extent than for auditory-vocal tasks.

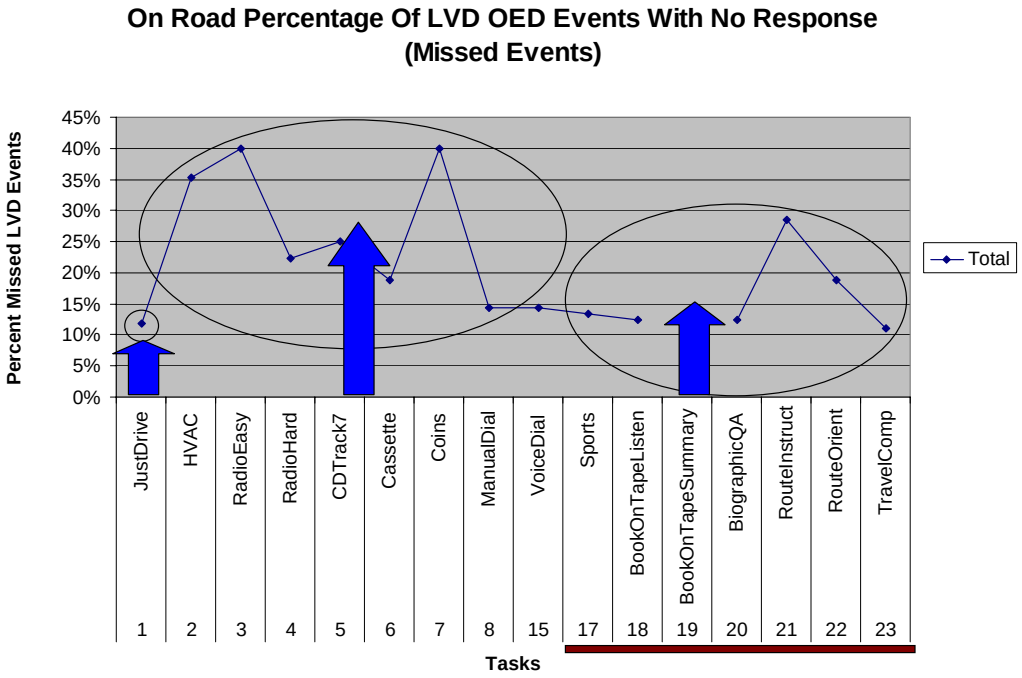


Figure 4-23. Percent LVD Missed on the Road as a Function of Tasks (for comparison to glance patterns)

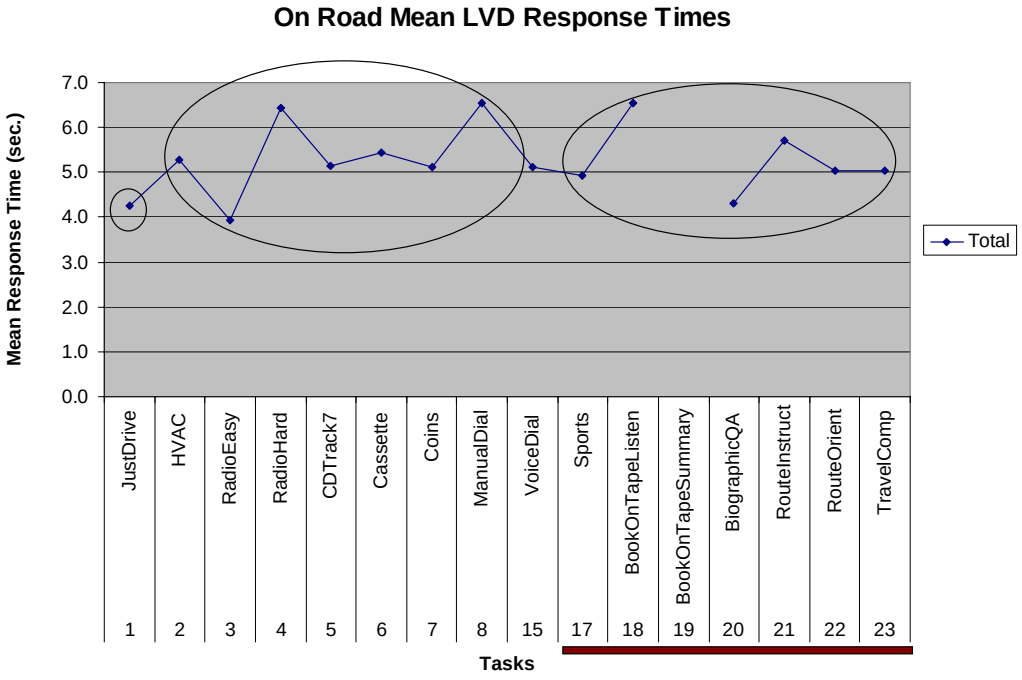


Figure 4-24. Response Times to LVDs Responded to on the Road as a Function of Tasks (for comparison to glance patterns)

In summary, the effects observed on the road for in-vehicle tasks on eyeglance behavior were about the same as those observed on the test track. Eyeglance metrics showed distinct patterns for different types of task engagement (just driving versus concurrently performing an auditory-vocal task or concurrently performing a visual-manual task). The Just Drive task was distinguished by patterns in which drivers looked at the road about 81 percent of the time and scanned their mirrors about 15.4 percent of the time. Glances on the road were about 3.8 seconds long, on average. Auditory-vocal tasks showed a somewhat similar pattern, though drivers gazed at the forward roadway somewhat more (87%), using longer gazes (4 to 8 seconds, on average), and scanned their mirror somewhat less (11.4%). The miss rate for event detection was slightly elevated over just driving for auditory-vocal tasks for CHMSL and LVD events (showing an increase of ~6% for CHMSL and ~4% for LVD events), and somewhat more for peripheral FVTS (an increase of ~14%)—though event detection was less affected by auditory-vocal tasks than by visual-manual tasks. Visual-manual tasks showed a different pattern, in which drivers looked at the forward roadway much less (viewing the road only 42 percent to 68 percent of the time during a task), and using glance durations on the road that were less than 2 seconds long, on average. This reduction in glances to the road was made in order to view task-related areas required for performing the in-vehicle activity (viewing the task 24 percent to 52 percent of the time during its length). For visual-manual tasks, glances tended to cycle frequently back-and-forth between the task and the roadway locations, and glance-rate measures proved to carry interesting information. Visual-manual tasks led to a more pronounced reduction in mirror-scanning (to 6.4%) and were associated with higher rates of missed events (though this was sometimes due to a methodological constraint for Lead Vehicle Decelerations). Increases in miss rates over Just Drive were approximately 23 percent for CHMSLs, 28 percent for LVDs, and 65 percent for FVTS events on average.

4.4.2 Event Detection and Eyeglance Patterns Relationships

Analyses of the eye data from the test track led to the finding that when drivers detected and responded to an event such as a CHMSL, FVTS, or LVD, their visual scan patterns changed. Therefore, analyses were undertaken on the road data to investigate whether or not such a relationship was borne out in more formal analyses of the glance data from the road as well.

Linear mixed-model analyses were conducted exactly like those done on the test track data, to provide further statistical tests of the hypothesis that events (like CHMSL illumination or LVDs) may function as attentional interrupts which serve to attract additional scanning which would increase situational awareness in relation to a possible threat or risk. These linear mixed-model analyses were separately conducted on each type of event detection (CHMSL, FVTS, and LVD). Emerging from these linear mixed-model analyses were significant main effects of Task, Location, and Detect Event (or Event Response), which were qualified by interactions of Location by Detection and, in some instances, Task by Detection. See Table 4-4.

Table 4-4. Linear Mixed-Model Effects For Analyses of CHMSL, FVTS, and LVD Detection Responses and Their Effects on Eyeglance Behavior

Road	Effect	# Glances	Total Dur	Max Dur	Min Dur	Mean Dur	Medn Dur	St Dev Dur	Glance Rt
CHMSL	Task	*	*	*	*	*	*	*	*
	Location	*	*	*	*	*	*	*	*
	Detect CHMSL								
	Task * Detect				*				.
	Locat* Detect	*	*	*		*	*		*
Road									
FVTS		# Glances	Total Dur	Max Dur	Min Dur	Mean Dur	Medn Dur	St Dev Dur	Glance Rt
	Task	*	*	*	*	*	*	*	*
	Location	*	*	*	*	*	*	*	*
	Detect FVTS			*		*	*	*	
	Task * Detect			.				*	
	Locat* Detect	*	*						*
Road									
LVD		# Glances	Total Dur	Max Dur	Min Dur	Mean Dur	Medn Dur	St Dev Dur	Glance Rt
	Task	*	*	*	*	*	*	*	*
	Location	*	*	*	*	*	*	*	*
	Detect LVD					*			
	Task * Detect				.		*		
	Locat* Detect	*							*

The linear mixed-model analyses confirmed that although there were significant main effects of Task, Location (road, SA, task, and NA), and Event Response (yes/no) (labeled "Detect" in Table 4-4), there were also significant interactions involving these variables. Of particular interest were the statistically significant interaction effects of Location by Detect, which will be referred to as Event Response, on multiple metrics, as indicated by the asterisks in the table above (asterisks designate effects significant at the $p < 0.05$ level; periods (.) indicate effects that were marginally significant at $p < 0.08$). Green highlighting identifies effects that have been explored graphically as well as statistically. Graphs are included in what follows for comparison of key effects to test track findings.

In brief, these graphs suggest that when an event occurs and is responded to, eyeglance behavior changes such that:

- For CHMSL events:
  - Durations of glances to the road increased, but not for awareness locations
  - Rate of glancing decreased slightly to road and task-related areas, and increased to situation awareness areas (mirrors)
- For LVD events:
  - Durations of glances to the road shortened
  - Rate of glancing decreased to road and task-related areas, and increased to situation awareness areas
- For FVTS events:
  - Durations of glances to the road decreased
  - Rate of glancing to road and task-related areas decreased, and increased to situation awareness areas

Changes to glance durations interacted with task type and were more pronounced for Just Drive and auditory-vocal tasks than for visual-manual tasks, which usually showed a different pattern.

The results presented in the following sections can also be interpreted in a different way. It is possible that the pattern of eyeglance behaviors generally do not reflect the impact of Object-and-Event Detection (OED) stimulus detection. Rather, it is the OED stimulus detection that reflects the general pattern of eyeglance behavior. A time-series analysis of each participant's eyeglances for each trial of a task's duration is needed to determine the prevalence of patterns to support either explanation. A time-series analysis is important in future work. A causal relationship in either direction is premature at this point. Therefore, the results reported here should be considered in light of both alternative interpretations.

Figure 4-25, Figure 4-26, and Figure 4-27, show the significant Location by Event Response interaction on the Number of Glances metric. As can be seen, there was a large increase in the number of glances to the road and situation awareness location types when any event was detected and responded to (but little or no increase in glances to task-related areas). There is no change in glances to the NA area, i.e., the category for eyeglances where vision was obstructed. Note that the pattern for LVD events is even more similar to that for CHMSLs and FVTS events on the road than it was on the test track where the increase in glances to task-related areas following response to the LVD event was somewhat larger, though still smaller than the increase to the road and situation awareness locations.

Each plotted point was obtained by averaging across all glances to a location (such as to the roadway, or to situation awareness locations, or to task locations) and across all tasks. Visual-manual tasks were shorter, and were associated with more missed events (or non-responses, plotted in blue on these graphs). Visual-manual tasks thus contributed more data to the blue points than did the auditory-vocal tasks. Auditory-vocal tasks contributed more data to the pink points than did the visual-manual tasks. There is, thus, the possibility that "type of task" also interacts with location and event response—and this relationship will be graphically depicted in subsequent figures.

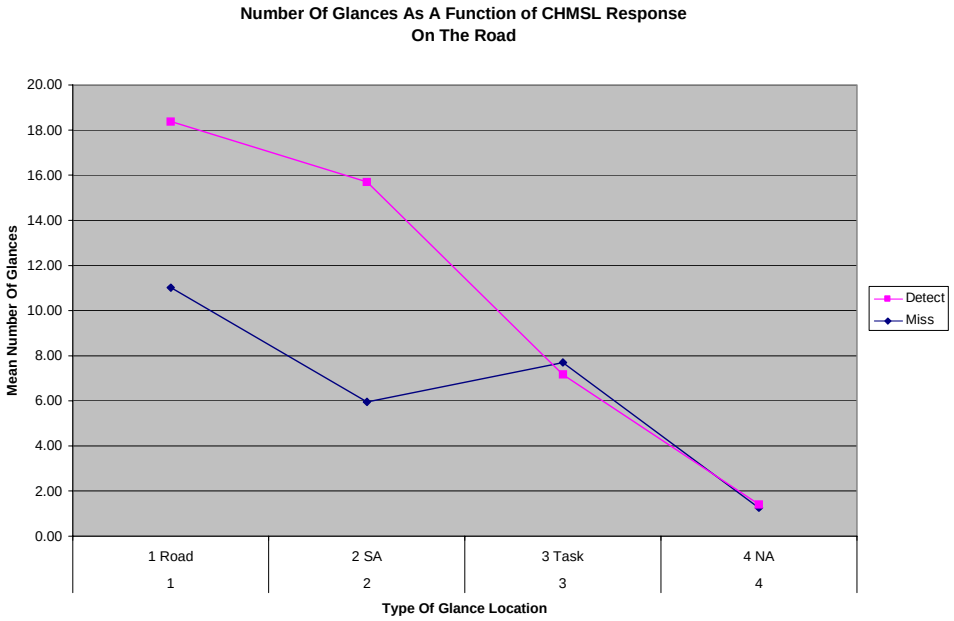


Figure 4-25. Effect of Response to CHMSLs on Number of Glances by Glance Location

Note: Shows a large increase in the number of glances to the road and SA location types when a CHMSL has been responded to (but not to task-related areas).

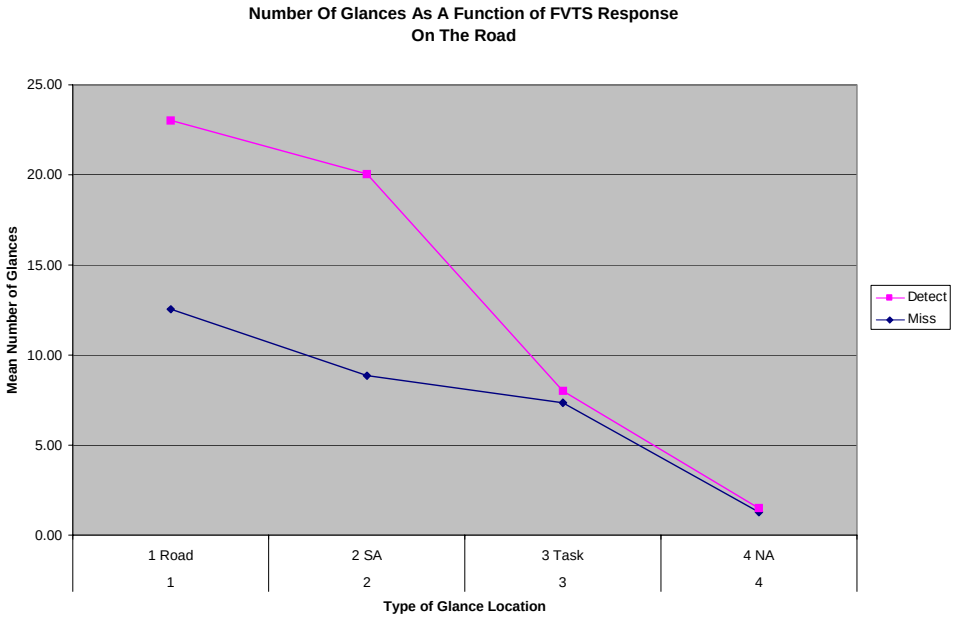


Figure 4-26. Effect of Response to FVTS on Number of Glances by Glance Location

Note: Shows a large increase in the number of glances to the road and SA location types when a FVTS has been responded to, but only a very small, almost negligible increase in glances to task-related areas

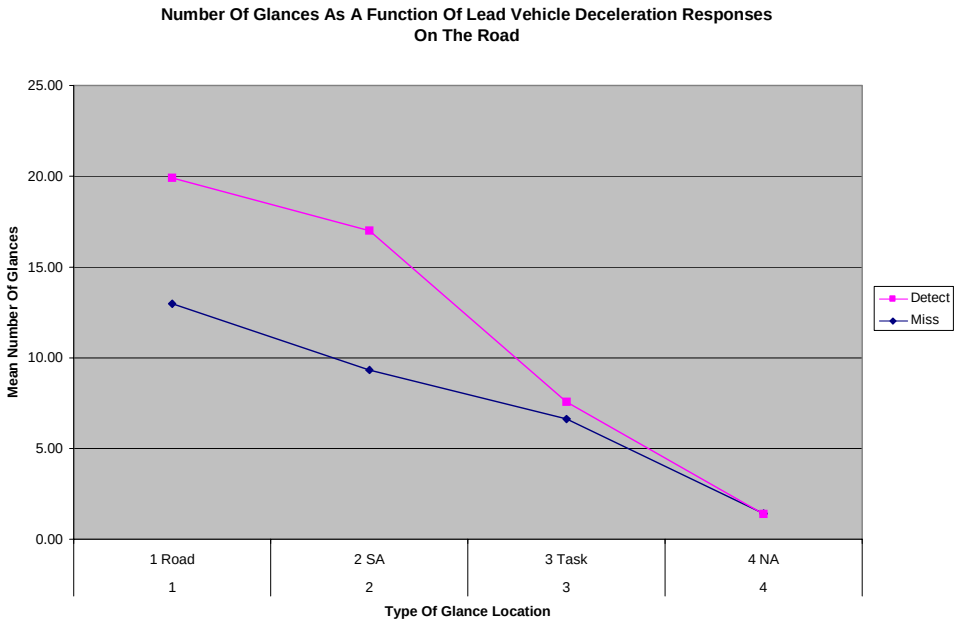


Figure 4-27. Effect of Response to LVD Responses on Number of Glances by Glance Location

Note: Shows a large increase in the number of glances to the road and SA location types, and also a smaller increase in glances to task-related areas

4.4.2.1 Glance Duration (as Affected by Location by Event Response)

Figure 4-28, Figure 4-29, and Figure 4-30 similarly show the interaction of Location by Event Response, but for metrics related to Glance Duration. The interaction indicates that, for CHMSLs, there is an increase in the duration of glances to the road (by slightly more than a second) following detection and response to the CHMSL, while the duration of glances to situation awareness and task-related areas remain relatively unchanged after detection and response to the CHMSL event. This contrasts with the test track finding of a small but reliable decrease in the duration of glances to all locations (road, task, and NA) except for those related to situation awareness (mirror and speedometer checks, which are already very short on average). The data for the road is shown in Figure 4-28, which depicts Mean Glance Duration, a metric on which this interaction was significant (as well as on Mean Glance Duration and Maximum Duration). For comparison, Figure 4-29 shows an opposite pattern for FVTS events, though it was not significant in the linear mixed models. Figure 4-30, shows the interaction on Median Duration for LVD events (also not significant, but provided for comparison with test track results). Unlike the test track results, it is similar in pattern to the interaction for FVTS, though the magnitude of change in road glance durations is smaller. Road glances shorten only by about a half second following detection and response to a LVD event. This contrasts sharply with the pattern found on the test track, where in the case of LVDs, glances to the Road area increased in duration rather than decreased. It was hypothesized that such a response would allow drivers to acquire more information about a decelerating vehicle over time, and so represents an appropriate adaptation to the detected lead vehicle deceleration. Conversely, longer glances to the road may have resulted in more LVD detections. Some evidence for this comes from task differences. Auditory-vocal and Just Drive tasks were generally much longer than Visual-manual tasks. Auditory vocal tasks contributed to a larger proportion of the trials averaged together for the “Detect” data point. On

the other hand, Visual-Manual tasks were shorter and had more "Miss" (missed detections) for perceptual reasons discussed later in this report. This alone could account for the differences in single-glance durations to the road.

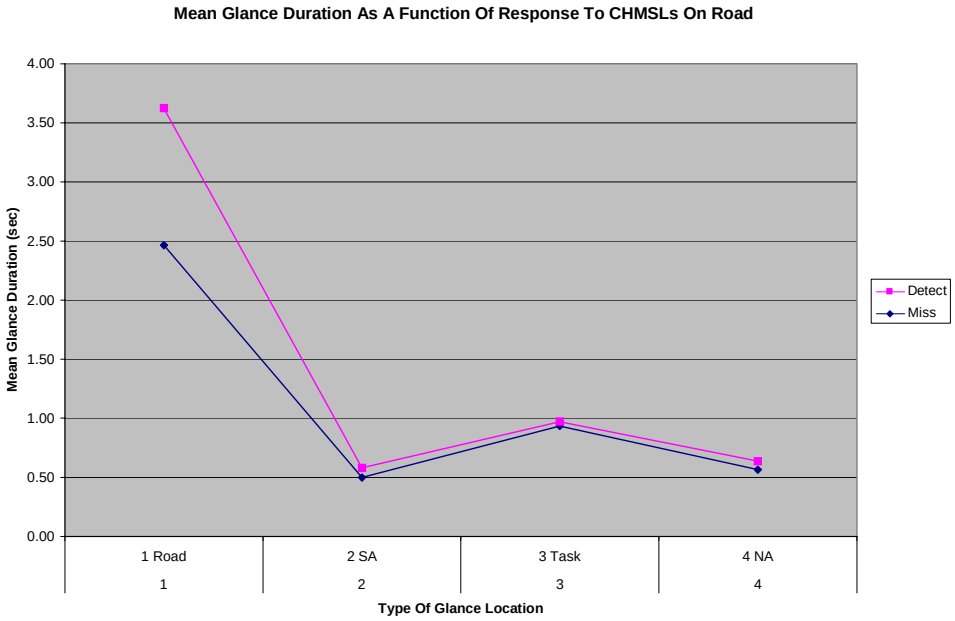


Figure 4-28. Road Mean Glance Duration by CHMSL Response Type and Glance Location

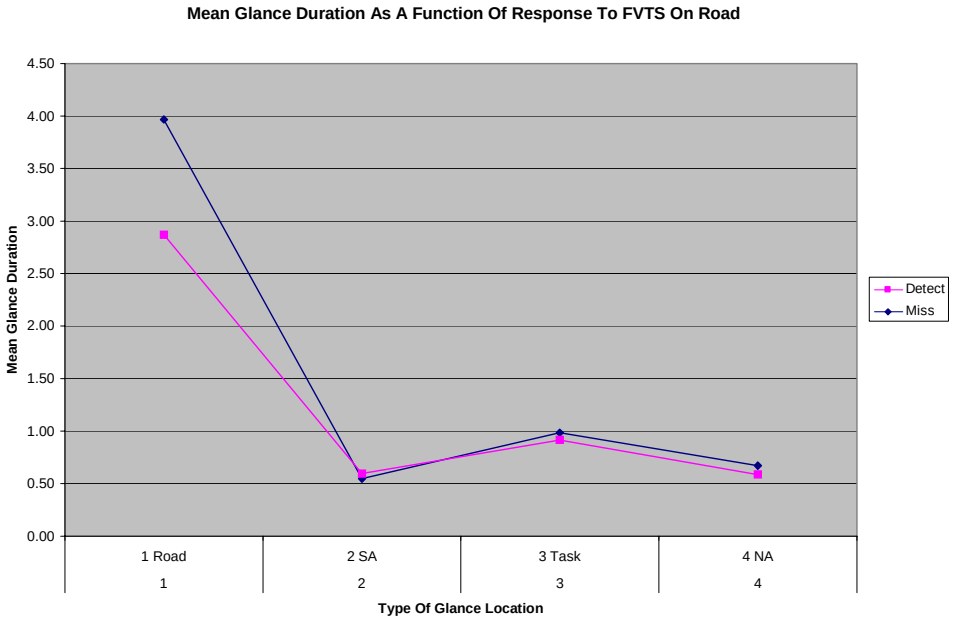


Figure 4-29. Road Mean Glance Duration by FVTS Response Type and Glance Location

Note: This Location by Event Response interaction was non-significant for FVTS events, though the pattern was consistent with CHMSL events.

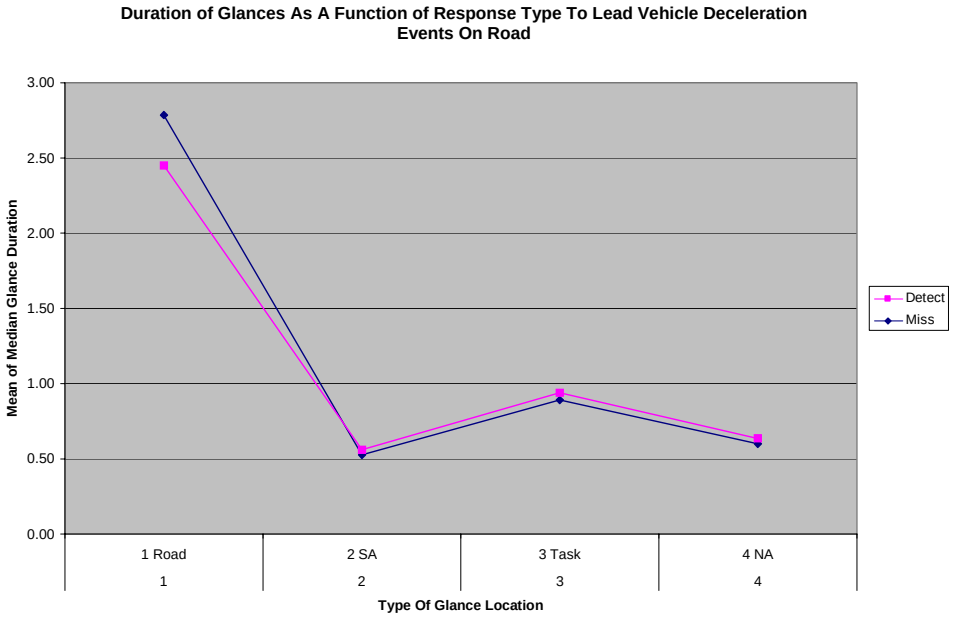


Figure 4-30. Mean Glance Duration by LVD Event Response Type and Glance Location

Note: This Location by Event Response interaction was non-significant for LVD events, though the pattern was consistent with CHMSL events.

4.4.2.2 Glance Rate (as Affected by Location by Event Response)

Figure 4-31, Figure 4-32 and Figure 4-33 show the significant Location by Event Response interactions for Glance Rate. As might be expected from the prior results, changes in number of glances and durations of glances following event detection and responses translate to changes in glances per second. Figure 4-31 depicts the interaction for CHMSL events, Figure 4-32 depicts it for FVTS events, and Figure 4-33 depicts it for LVD Events. All three figures indicate a decrease in glances-per-second to the road and task-related locations, and an increase in glances-per-second to the situation awareness location (mirrors/speedometer). The decreases in glance rate to the road and task was largest following response to a CHMSL event. The increases in glance rate to the situation awareness location were largest following detection of the CHMSL and FVTS events. But, on the road, responses to an event of any type led to similar changes in glance rates with the adaptations varying by event-type and location only in terms of magnitude.

These results again differ somewhat from the test track results for the same interactions. On the test track, the results for the LVD events bear some resemblance to those for the road. For CHMSL and FVTS, glance rates to the road increased, and glance rates to the task increased or remained relatively stable. Even so, the road data are consistent with the notion that driver glance patterns are modified in response to events that are detected and responded to during driving, and that they are modified in a way that is specific to the event. The results suggest that drivers adapt their visual scanning in a way that is perhaps tailored for updating their awareness of current traffic and road conditions relative to the specific event they are responding to, and the types of risks it may represent to them within the context of the driving conditions at the moment. For example, as mentioned previously, when a CHMSL illuminates in front of them, drivers may habitually check mirrors to determine whether a lane change may be possible should the vehicle in front suddenly stop. When a follow vehicle signals a turn, looks to the mirror may increase to

ascertain whether an overtaking maneuver will be initiated. And even though these conditions were not really relevant under the experimental conditions of the platoon methodology used in this experiment, scan patterns learned over years of driving may nonetheless be triggered by the stimulus events used in the study.

As mentioned in the test track discussion of results, however, it may be that the effects observed here were in some way unique to the event detection methodology employed in the experiment. For example, while a driver may habitually check mirrors in case a sudden stop by a lead vehicle requires an evasive lane change maneuver, the likelihood of this was reduced on the test track by (1) extremely light traffic, and (2) no sudden hard braking for task after task, which might have been expected to cause drivers to learn that hard braking was unlikely, purportedly a driver-expectation that is a common contributor to rear-end crashes. If drivers in the study in fact did not expect hard braking, and were not changing their scanning patterns due to learned responses that are adaptive for driving, then perhaps they changed their scanning merely to detect events that they expected in the experimental paradigm. However, this explanation cannot account for a change in glance patterns on the test track after the detection of an event, since only one event per task was presented for detection in that venue. Thus, on the test track, there would have been no point to changes in glance durations or increased scanning of road and mirror locations following detection of an event for experimental purposes, since it would not have improved event detection performance during the task. In contrast to the test track, on the road, multiple events (up to one of each type) were presented during the long (auditory-vocal) tasks, so it is possible that such a strategy may have partially been at play in the on-road results.

For the road venue, both explanations have some plausibility. Therefore, further studies to determine why such shifts in scan patterns may occur as a function of event detection would be desirable.

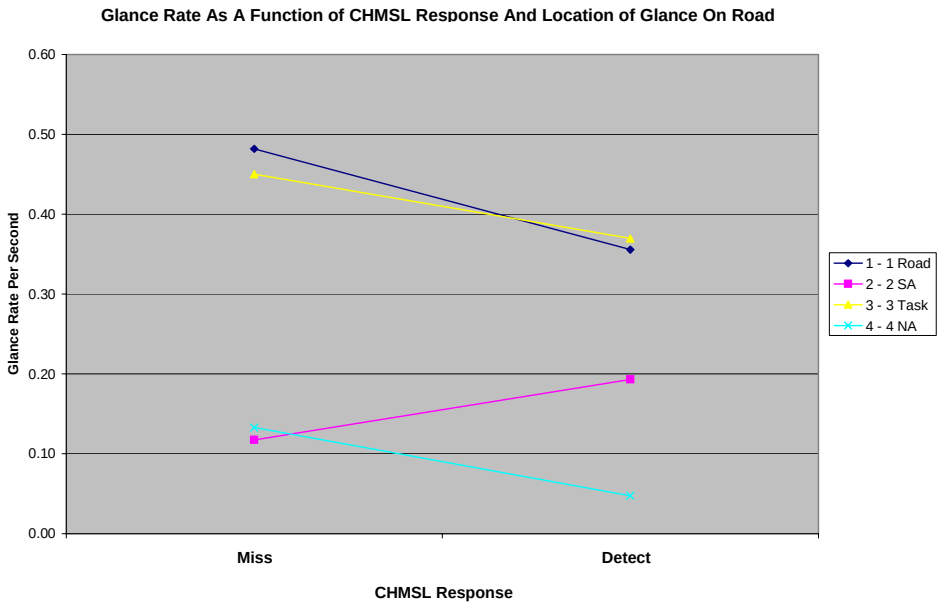


Figure 4-31. Road Glance Rate Metric by Event Response for CHMSL Events and Glance Location

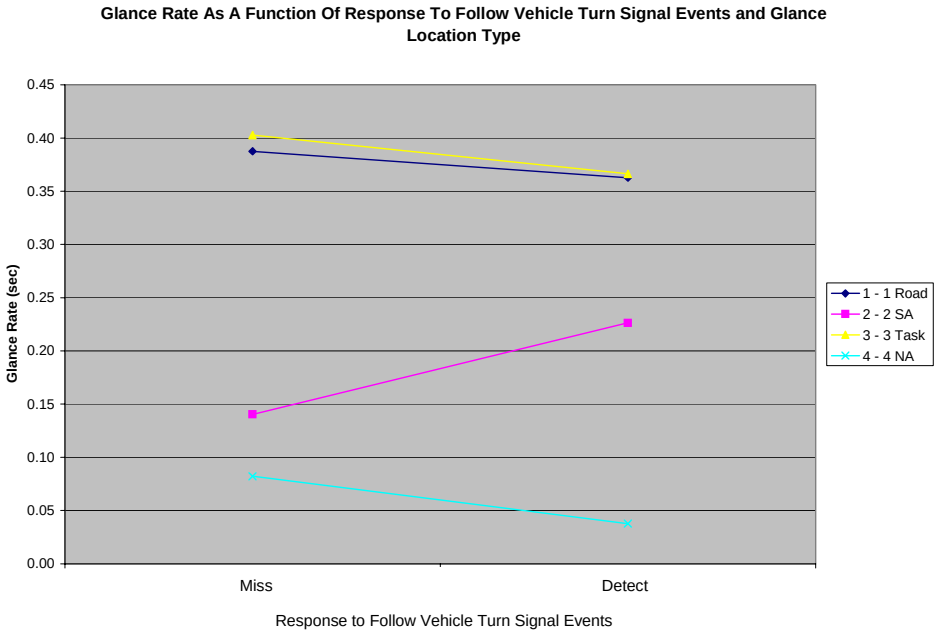


Figure 4-32. Road Glance Rate Metric by Event Response for FVTS Events and Glance Location

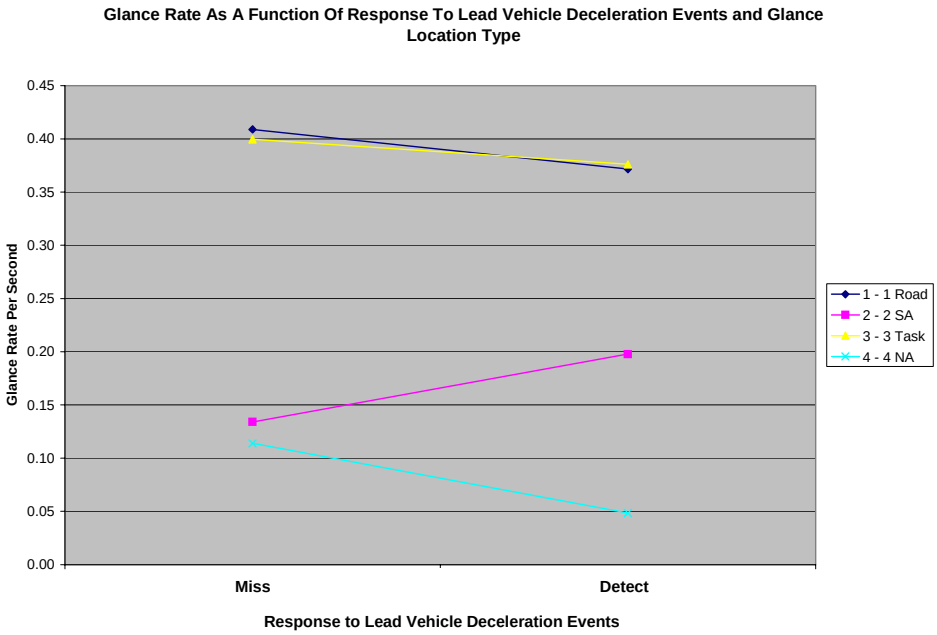


Figure 4-33. Road Glance Rate Metric by Event Response for LVD Events and Glance Location

4.4.2.3 Glance Duration (as Affected by Task by Event Response)

Figure 4-34, Figure 4-35, Figure 4-36, and Figure 4-37 show the interaction of Task by Event Response on the metric of Glance Duration (averaged across all location types). In these figures, for each task, each blue or pink point was obtained by averaging across all glances to all locations during that task. Further, because the set of data is based upon only those 18 participants from

whom eye data were reduced form the test track venue, when these data are decomposed by task and then in terms of whether an event was detected or not, the data are sparse in some cells (for example, there are some cells in which no missed event-detections occurred). In those instances, a point will be missing from the plots below.

The interaction for CHMSL events is illustrated in Figure 4-34, using Mean Glance Duration, for which it was not significant, to enable comparison with test track results and Figure 4-35, Maximum Glance Duration, for which it also was not significant. In the road data, it was significant only for Minimum Duration. In this interaction, it can be seen that the decrease in glance durations following detection and response to the CHMSL events is confined to only two tasks, primarily Book-on-Tape Summarize, and Route Orientation. This is due in part to the fact that there are no comparison points for some tasks, given that all CHMSLs were detected and responded to for some tasks, and hence there are no "Miss" points plotted for them. Note that the majority of these tasks are auditory-vocal tasks, which are typically characterized by long glances at the road during task performance, and these glances shorten following response to a CHMSL event and thus, there would be more of them. However, the interaction for CHMSLs is also due to the fact that for other tasks-most of the visual-manual tasks and two auditory-vocal tasks-the pattern is different. For the visual-manual tasks, there is virtually no change in Mean or Maximum Glance Duration as a function of having detected and responded to the CHMSL event. For the two auditory-vocal tasks, Route Instructions and Travel Computations, and the mixed-mode task of Voice Dial, there was an increase in Maximum Glance Duration (across all location types). However, even though glance durations shorten for most of the auditory-vocal tasks following detection of a CHMSL event, they remained longer than for visual-manual tasks (by a factor of three or more in most cases). As mentioned previously, this was likely due to the fact that most of the glances for auditory-vocal tasks were to the road location and longer versus split between the task and road and, hence, shorter.

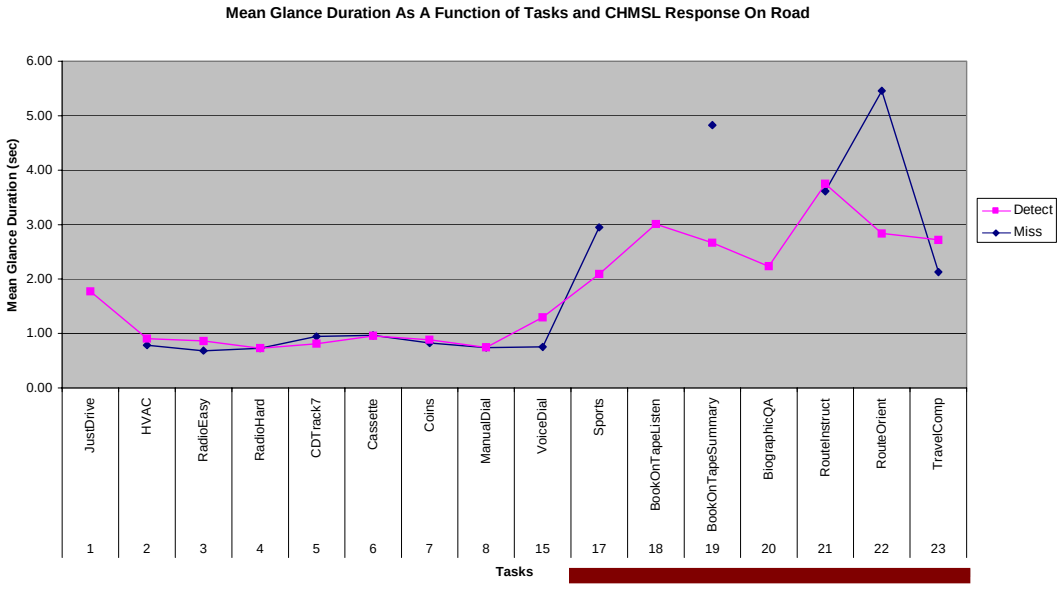


Figure 4-34. Non-significant Task by Event Response Interaction for CHMSL Events on the Metric of Mean of Mean Glance Durations (shown for comparison with other patterns)

Note: There were no missed detections of CHMSL for some tasks.

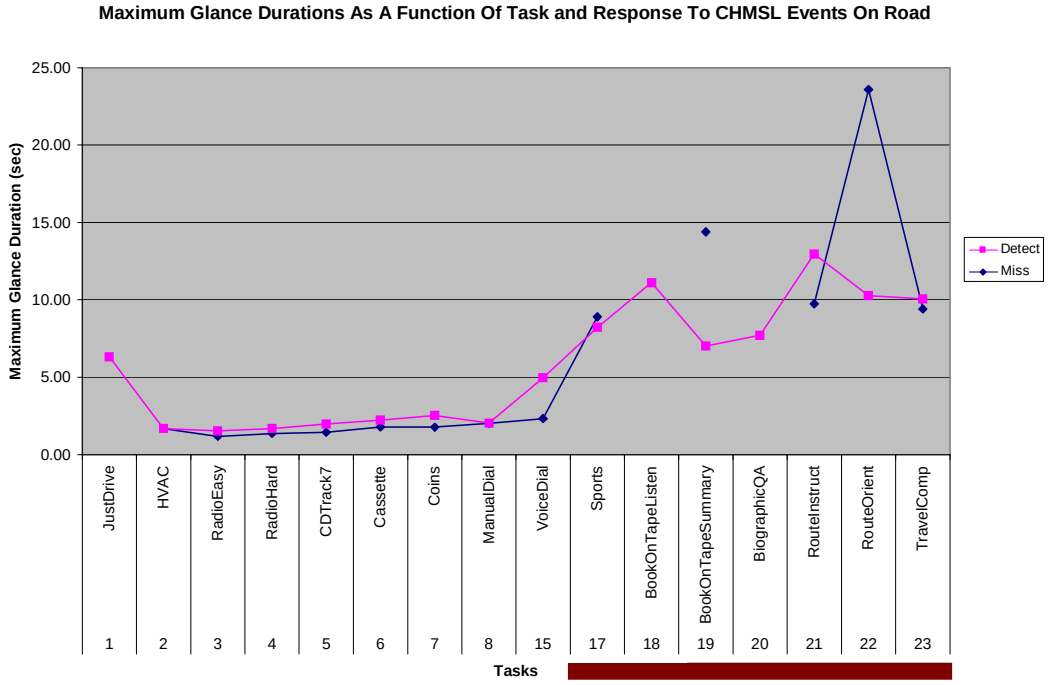


Figure 4-35. Marginally Significant Task by Event Response Interaction for FVTS Events on the Metric of Mean of Maximum Glance Durations (shown for comparison with test track results)

The interaction for the FVTS events is shown for the road data in Figure 4-36. This figure also shows that the decrease in Mean Glance Duration across all location types, which is significant as shown in Table 4-4, is confined to a small subset of tasks (Just Drive and the auditory vocal tasks). Visual-manual tasks show little or no change in mean glance duration as a function of FVTS event detection and response. Interestingly, following the detection of an FVTS event, the mean glance duration (again, averaged across all locations) for Just Drive and auditory-vocal tasks more closely resembles that for visual-manual tasks, but is still distinctly longer on average. This figure for the road data (as was true for the corresponding figure for test track data) shows that this effect is not due just to task duration. Book-on-Tape Summarize was a short auditory-vocal task, only about 35 seconds in duration versus approximately two minutes for the others, and still demonstrated the drop in glance durations for trials on which an event was detected.

Figure 4-37 depicts the interaction of Task by Event Response for LVD Responses on the metric of Mean Glance Duration. On this metric, the interaction was not significant, although it was on Minimum Duration and Median Duration for the road data. However, it is shown to allow comparison with the test track data. It also indicates that the changes to glance duration primarily occurred on a very small subset of tasks, which were auditory-vocal in nature (Biographical Q&A, Route Orientation, and Travel Computations), along with Just Drive. On the test track, lengthening of glance durations was observed for some tasks, and was consistent with some of the results from the test track. The pattern for LVD showed that glance durations lengthened on the auditory-vocal tasks of Sports Broadcast (very slightly), Book-on-Tape Listen, and Route Instructions. Visual-manual tasks showed little change in glance durations.

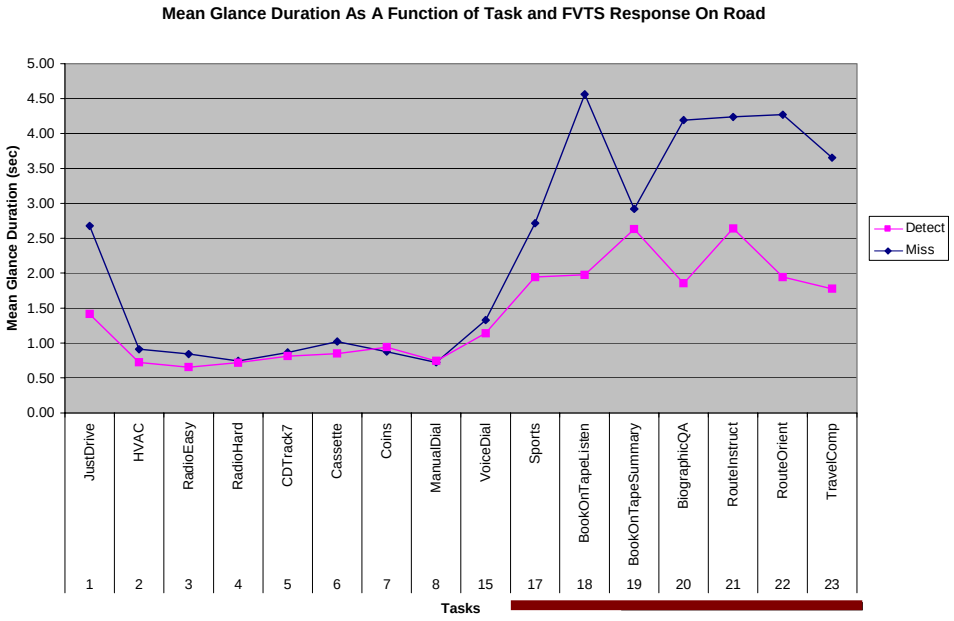


Figure 4-36. Non-Significant Task by Event Response Interaction for FVTS Events on the Metric of Mean of Mean Glance Durations (provided for comparison with test track results).

Note: The interaction was significant for Standard Deviation of Glance Durations in the road data.

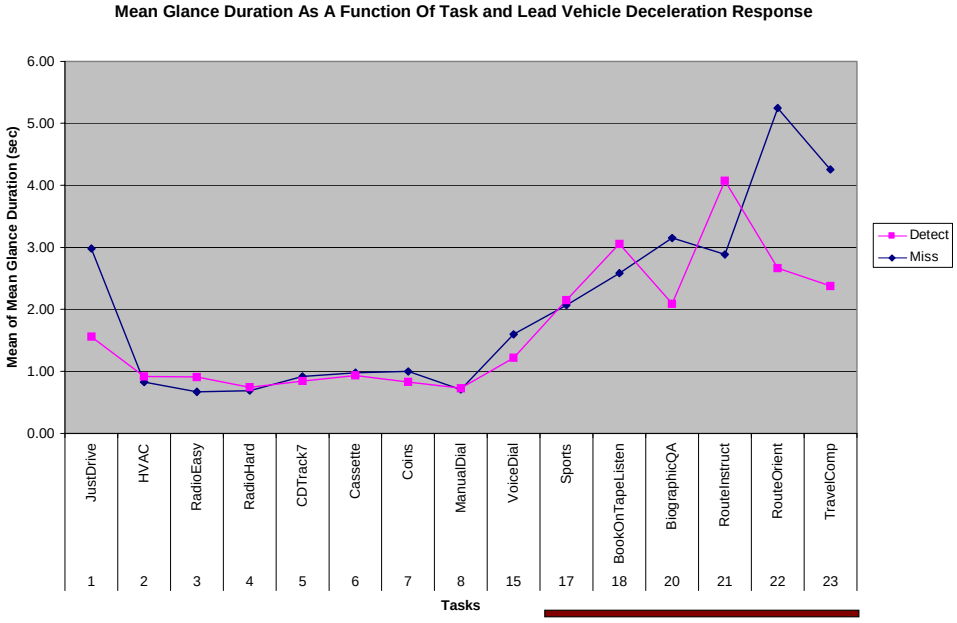


Figure 4-37. Significant Task by Event Response Interaction for LVD Events on the Metric of Mean of Mean Glance Durations (provided for comparison with test track results)

Note: The interaction was significant for Minimum Duration and Median Duration in the road data.

4.4.2.4 Glance Rate (as Affected by Task by Event Response)

Figure 4-38, Figure 4-39 and Figure 4-40 show the significant interaction of Task by Event Response for the road data on the Glance Rate metric for each of three event types. The pattern for the road data was different from the pattern for the test track data. For CHMSL events, shown in Figure 4-38, glance rates averaged across all location types dropped from some tasks (e.g., Just Drive, HVAC, Radio (Easy), Radio (Hard)) remained the same or similar for some tasks (e.g., CD/Track 7, Coins, etc), and may have increased for others (e.g., Book-on-Tape Summarize, Route Instructions, Route Orientation). This contrasted with the test track finding in which glance rates averaged across all location types increased for nearly all tasks following detection of a CHMSL, with a few exceptions (HVAC, Map (Hard), Route Tracing, and Book-on-Tape Listen). This appears to be due in part to the fact that the glance rates on trials where the CHMSLs were not responded to were already very high on the road (higher than on the test track), so a careful consideration of the magnitudes of the glance rates between road and track on a task-by-task basis deserves further investigation. It may be, for example, that the workload associated with the road-and-traffic environment on the road was higher and caused drivers to adopt different strategies of glancing during task performance, including different rates of glancing.

The interaction for FVTS events is shown in Figure 4-39. Again, the pattern for the road data is somewhat different from the pattern for the test track. On the test track, the glance rates increased following detection of the FVTS event for all tasks, though the increase for Route Tracing was negligible, and the increases for visual-manual tasks tended to be smaller than for auditory-vocal tasks. For the road data, increases in glance rates were observed only for Just Drive and the auditory-vocal tasks. For the visual-manual tasks, the glance rates following a response to an FVTS were similar to those observed on the test track. However, the glance rates following a non-response or missed detection of an FVTS were higher on the road than on the test track, as if the drivers were in a higher scan state for some other reason on these trials. This finding again deserves more investigation.

The interaction for LVD events is shown in Figure 4-40 and showed some elements of consistency with the results reported for the test track on Glance Duration and Glance Rates, as well as for this interaction. It shows that Glance Rate decreased for some of the visual-manual tasks (e.g., Radio (Easy), Radio (Hard), CD/Track 7) but increased for others (Insert Cassette, Coins) and increased for some auditory-vocal tasks (Sports Broadcast, Biographical Q&A, Route Orientation, and Travel Computations), Just Drive, and the mixed-mode task of Voice Dial. As discussed in Chapter 3 for visual-manual tasks, a reduction in glance rate for visual-manual tasks would be consistent with some type of reduced scanning between task and roadway locations, in order to attend to the LVD event. This same reduction was not seen for CHMSL and FVTS events, however. For auditory-vocal tasks, though, it is the opposite pattern that would indicate a shift of attention to event monitoring. Namely, an increase in glance rate would indicate that a shift from steady gazing at the road to active scanning of the forward roadway and mirrors during auditory-vocal tasks. This is, in fact, what occurred.

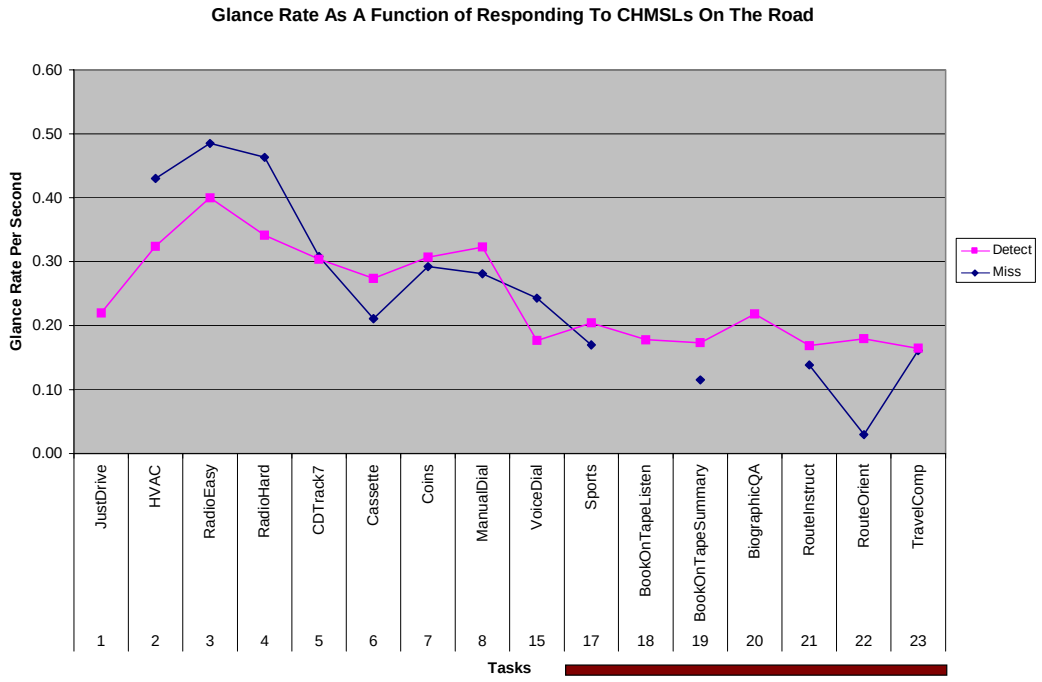


Figure 4-38. Task by Event Response Interaction for CHMSL Events on the Glance Rate Metric

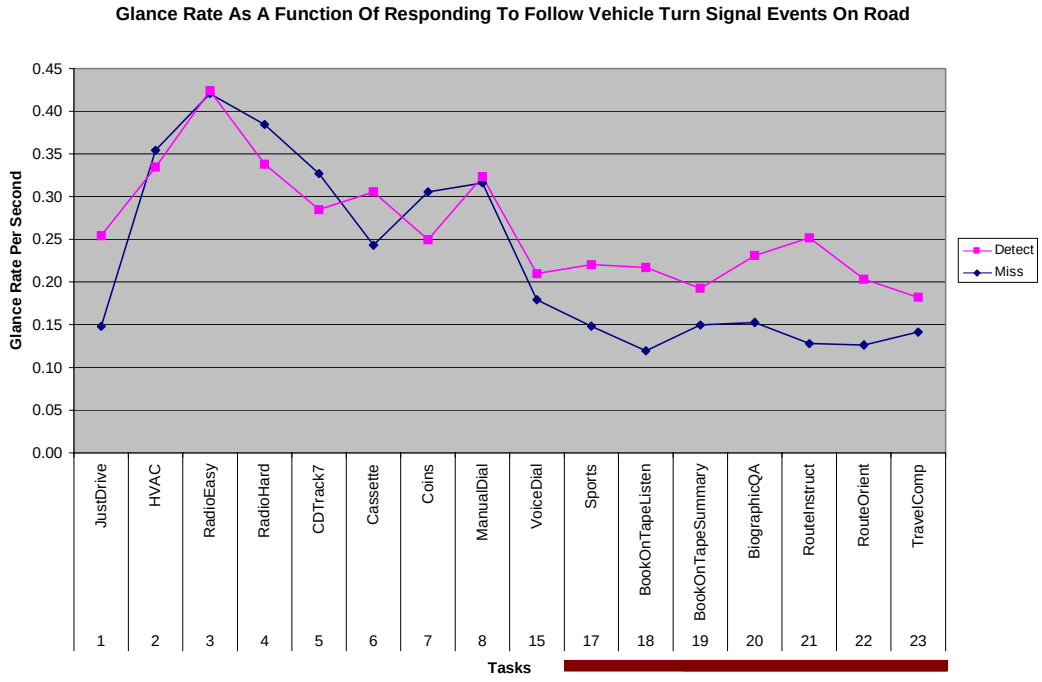


Figure 4-39. Task by Event Response Interaction for FVTS Events on the Glance Rate Metric

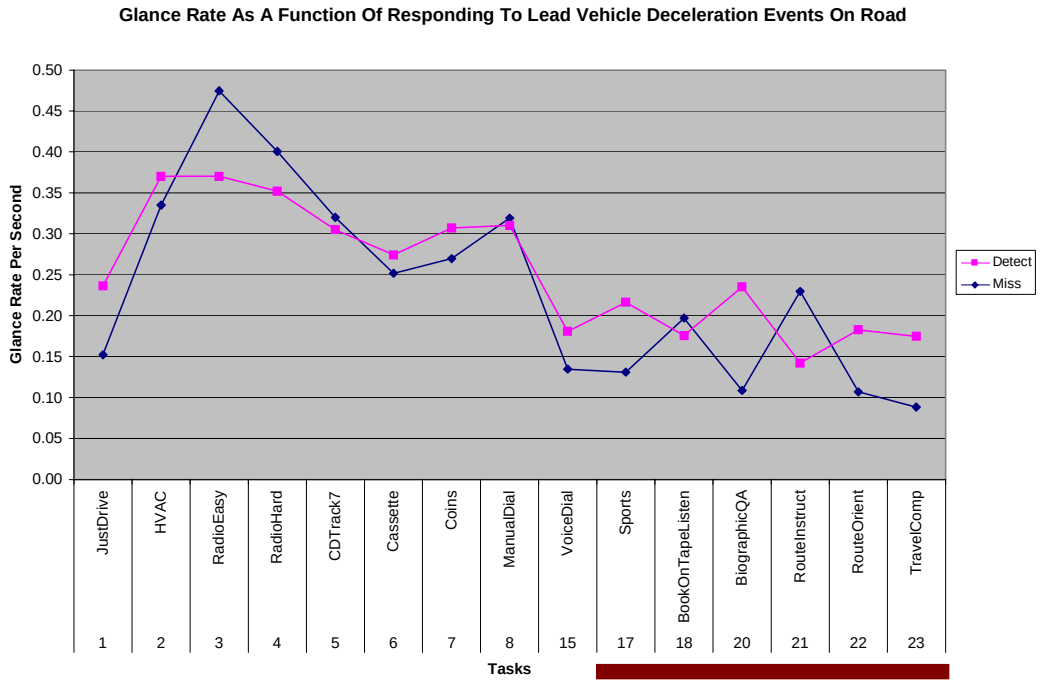


Figure 4-40. Task by Event Response Interaction for LVD Events on the Glance Rate Metric

To summarize, the results of formal analyses of the on-road effects of event detection on glance patterns, the main finding from the test track was replicated. When events were detected and responded to by drivers, scan patterns subsequently changed in a way that appeared adaptive to the specific type of event to which the driver had responded. However, the exact changes observed on the road differed in some ways from the changes observed on the track.

As discussed previously, the finding that event detection affects glance patterns is one that has methodological, theoretical, and practical importance and deserves considerable further study in future work. These implications are recapped below.

Theoretically, the implications of these findings are that event-detection may serve as an "attentional interrupt" for auditory-vocal tasks and the task of just driving, resulting in more active scanning of the road and mirrors for situational awareness. The same phenomenon may hold for certain types of OED stimulus events with visual-manual tasks as well. For visual-manual tasks, this appeared to occur only for LVD events. For other event types, when an event was detected during a visual-manual task, visual scanning between task, road, and mirror locations seemed to increase (apparently without task shedding), and high glance rates appeared to be related to higher miss rates for these events. Hypotheses about the effects of events on the deployment of attention during driving need to be developed and confirmed in further work, particularly work that is done in a more naturalistic setting to see if event detection effects that were observed here were due to the experimental paradigm or conditions used, or whether they will generalize to naturalistic driving. The road results indicated that traffic conditions had a further effect on the ways in which glance patterns changed (perhaps an effect of loading and/or context), and may suggest that the underlying processes through which the driver determines where to glance next and for how long may be especially important to understand. Some of these processes may occur outside of conscious awareness, and some may occur within conscious

awareness. As such, both brain imaging and behavioral science approaches may be needed in order to push the state of understanding forward in this arena.

From the point of view of methodology and practicality, there are several implications of event detection effects, if confirmed in further research, for measuring glance behavior in evaluations of advanced information systems. Eyeglance behavior collected during auditory-vocal tasks or Just Drive may need to involve trials with and trials without OED stimulus events. Visual-manual tasks may not need multiple sets of trials, as long as slowly evolving events are not used as detection stimuli. When evaluating the visual demand of tasks in an advanced information system or in-vehicle device, it may be important that multiple test trials be conducted, some with and some without event detection. How important this is appears to depend on the type of task and the type of OED stimulus. Under certain conditions indicated above, the trials used to evaluate the visual demand of a task should not include events to-be-detected. Because the presence of events-to-be-detected can change durations and numbers of glances, depending on the type of event that is presented, these events can spuriously alter the visual demand assessment results if they are included in trials used to assess visual demand. Ideally, an assessment would be done in a context in which drivers sometimes received visual events during tasks, and sometimes did not. The drivers would not know on which trials events would occur, and so would have to be monitoring for events on each trial. However, on the test trials actually used to assess visual demand for a task, no event would be presented. These trials would yield clean measurements of glance behavior, free from influence of co-occurring events.

4.4.3 Analyses of Reliability and Predictive Validity for Glance Metrics

In evaluating the properties of the measures taken on the road, analyses of reliability and predictive validity for the eyeglance measures were undertaken in a manner consistent with those done on other categories of measurement.

4.4.3.1 Overall Level – Split-Half Repeatability

To examine the reliability of eyeglance measures, the sample of data collected on the road was split in half and correlations between the split halves were computed, following the methods previously described. However, it should again be noted that eye data from the track came from only 18 research participants. This meant that the split halves were well-balanced for the data analyses, but not perfectly balanced (with nine in each subset) by age and gender, as they were for most other subsets of data. Nonetheless, there was a desire for all split-half analyses to be similarly implemented, so the same assignments of research participants to split halves were used for the eye data analyses as were used in all other analyses of Split-Half Repeatability.

The split-half correlations were done across the full set of tasks performed on the road. The outcome of the correlations between the split halves is shown in Table 4-5. The second column shows the split-half correlations for the road data, and the third column shows the correlations from the test track data (for comparison). The extended variable names in the rightmost column of Table 4-5 are shown and highlighted for those variables that proved reliable across both test track and road venues. Those items that are highlighted in green in the table had correlations greater than + 0.707. Items with correlations greater than 0.665 (still significant at $p < 0.05$) are highlighted in a softer shade. Generally, the eyeglance measures, which met this criterion for repeatability, fell into a small number of groups. However, it is possible to see by comparing the first column (road data) to the second column (test track data) that more metrics were repeatable in the road data than in the track data.

Metrics related to the following categories were repeatable:

- Number of glances
  - To road locations
  - To situation awareness locations
  - To task-related areas
  - To total/all, and to not road (combines everything other than road)
- Durations of glances
  - Mean (for most locations – road, situation awareness, task (was borderline), and not road)
  - Median (for some locations – road, situation awareness)
  - Standard deviation (for some locations – road, task, total/all, not road))
  - Max (for only certain location types – road, task, total/all)
- Accumulations of durations
  - Total Glance Time to Road Location
  - Total Glance Time to Situation Awareness Location
  - Total Glance Time to Task-Related Areas
- Percents (or proportions) of task time spent looking at a location type
  - To road locations
  - To situation awareness areas (borderline)
  - To task-related areas
- Rates of glances per second
  - Overall (total/all), road, task, not road)

Table 4-5. Split-Half Correlations for On-Road Data on Eyeglance Measures

Eye Glance Metric	ROAD DATA Split-Half Reliability, Pearson r	TRACK DATA Split-Half Reliability, Pearson r	Eye Glance Metrics Repeatable Across Both Venues
MeanTskglncs	0.947	0.852	**Total Glances to Any Location During Task**
MeanTaskdur	0.994	0.996	**Mean Task Duration Derived From Eye Data**
MeanmeanTdur	0.870	0.681	**Mean of Mean Duration Of All Glances (To All Locations) During Task**
MeanmedTdur	0.382	0.518	
MeansdTdur	0.837	0.829	**Mean Stand. Deviation of Task Duration**
MeanTglsprs	0.974	0.930	**Mean Rate of Glances Per Second During Task** (includes glances to all locations)
MeanglncesRD	0.949	0.852	**Mean Number of Glances To Road**
MeanduratRD	0.995	0.996	**Mean Total Glance TimeTo Road During Task (Summed Across Glances)**
MeanmeanRDdr	0.886	0.811	**Mean of Mean Glance Durations To Road**
MeanmedRDdur	0.834	0.788	**Mean Of Median Glance Durations To Road**
MeansdRDdur	0.816	0.839	**Mean Stand. Deviation of Glance Durations To Road**
MeangrateRD	0.973	0.917	**Mean Glance Rate Per Second To Road (During Task)**
MeanpctdurRD	0.992	0.983	**Mean Percent of Task Duration Spent Looking At Road**
MeanglncesSA	0.983	0.951	**Mean Number of Glances To Sit Awareness Locations (Mirrors & Speedo)**
MeanduratSA	0.983	0.976	**Mean Total Glance Time To Sit Awareness Locations (Summed Across Glances)**
MeanmeanSAdr	0.850	0.828	**Mean of Mean Glance Durations To Sit Awareness Locations**
MeanmedSAdur	0.797	0.781	**Mean of Median Glance Durations To Sit Awareness Locations**
MeansdSAdur	0.497	0.548	
MeangrateSA	0.750	0.245	**Mean Glance Rate Per Second To Sit Awareness Locations (Mirrors & Speedo)**
MeanpctdurSA	0.884	0.675	**Mean Percent Duration of Task Spent Looking At Sit Awareness Locations**
MeanglncesTR	0.934	0.983	**Mean Number of Glances To Task-Related Areas**
MeanduratTR	0.960	0.977	**Mean Total Glance Time to Task-Related Areas (Summed Across Glances)**
MeanmeanTRdr	0.733	0.677	**Mean of Mean Duration of Glances To Task-Related Areas**
MeanmedTRdur	0.871	0.599	**Mean of Median Duration of Glances To Task-Related Areas**
MeansdTRdur	0.459	0.784	**Mean Stand. Deviation of Glance Durations To Task-Related Areas**
MeangrateTR	0.955	0.932	**Mean Glance Rate Per Second To Task-Related Areas**
MeanpctdurTR	0.993	0.939	**Mean Percent Of Task Duration Spent Looking At Task Locations**
MeanglncesNA	0.035	0.147	
MeanduratNA	0.021	0.661	
MeanmeanNAdr	-0.238	0.497	
MeanmedNAdur	-0.241	0.455	
MeansdNAdur	0.740	-0.112	
MeangrateNA	0.699	0.904	**Mean Glance Rate Per Second During Task Spent Looking At N.A./Obstructed**
MeanpctdurNA	0.960	0.940	**Mean Percent Duration of Task Spent Looking At NA Locations**
MeanglncesMR	0.982	0.961	**Mean Number of Glances At Mirrors Alone**
MeanduratMR	0.983	0.984	**Mean Total Glance Time To Mirrors Alone (Summed Across Glances)**
MeanmeanMRdr	0.876	0.750	**Mean of Mean Glance Durations To Mirrors**
MeanmedMRdur	0.789	0.675	**Mean of Median Glance Durations To Mirrors**
MeansdMRdur	0.508	0.492	
MeangrateMR	0.626	0.077	
MeanpctdurMR	0.832	0.558	**Mean Percent of Task Duration Spent Looking At Mirrors**
MeanglncesNR	0.936	0.851	**Mean Number of Glances To All Areas Classified As "NOT ROAD"**
MeanduratNR	0.885	0.919	**Mean Total Glance Time To All "NOT ROAD" Areas (summed across glances)**
MeanmeanNRdr	0.911	0.728	**Mean of Mean Duration Of Glances To "NOT ROAD" Areas**
MeanmedNRdur	0.897	0.633	**Mean of Median Duration of Glances to "NOT ROAD" Areas**
MeansdNRdur	0.842	0.896	**Mean Stand. Deviation of Glance Durations To "NOT ROAD" Areas**
MeangrateNR	0.961	0.941	**Mean Glance Rate Per Second To "NOT ROAD" Areas**
MeanpctdurNR	0.990	0.980	**Mean Percent Time During Task Spent Viewing "NOT ROAD" Areas**
MinTdur	-0.039	-0.145	
MinRDdur	0.039	-0.037	
MinSAdur	0.489	0.154	
MinTRdur	0.477	0.542	
MinNAdur	0.375	0.262	
MinMRdur	0.565	0.004	
MinNRdur	0.242	0.178	
MaxTdur	0.778	0.732	**Maximum Duration of Glances To All Locations During Task**
MaxRDdur	0.779	0.777	**Maximum Duration of Glances To The Road**
MaxSAdur	0.820	0.430	**Maximum Duration of Glances to Sit. Awareness Locations**
MaxTRdur	0.823	0.892	**Maximum Duration of Glances To The Task**
MaxNAdur	0.645	0.393	**Maximum Duration of Glances In The NA (Obstructed/Not Scorable) Category**
MaxMRdur	0.823	0.483	
MaxNRdur	0.344	0.367	

Items in green have "r" values >.707 (original cutoff for repeatability)

Items in blue have "r" values >0.665 and p<0.05

4.4.3.2 Predictive Validity – Correlations between Road Eyeglance Data and Road Driving and OED Performance Metrics

Measures of eyeglance behavior are considered fundamental to driving performance. Therefore, it is not necessary to establish whether they have predictive validity or whether other driving performance measures can be predicted from eyeglance measures. It is nonetheless informative to explore the relationships that exist between eyeglance metrics and other driving performance measures.

Correlations for the Full Set of Tasks (Between Eyeglance Data and Performance Data)

Table 4-6 presents the correlations between the eyeglance metrics and the reliable driving performance metrics across the full set of tasks (visual-manual, auditory-vocal, mixed-mode) for the road data. Results discussed in this section will subsequently be broken out by task type because different task types have different properties. For example, visual-manual tasks are associated with back-and-forth glance pattern between task and road. Thus, more glances to the road are accompanied by more glances to the task as well. Auditory-vocal tasks do not have this property. Median standard deviation of lane position (SDLP) correlated positively with a variety of eyeglance metrics. These included the number of glances made to any location throughout a task as a whole, mean number of glances made to the road, and mean total glance time to the road, as well as maximum glance duration to the Road and maximum glance duration of any type to any location. In addition, median SDLP was positively correlated with the metrics related to number and durations of glances to situation awareness locations (mirror and speedometer) and to just mirror locations (MR).

The more glances to the road and mirrors, and the longer these glances were, the higher the SDLP. There was also a positive correlation with Mean Task Duration (based on glance information). These relationships indicated that as task duration increased, number of glances increased to the road and mirrors increased, and so did median SDLP. These relationships are rather difficult to interpret and seem counterintuitive, since it would seem that more time spent looking at the road would lead to better (less variable) lane position. It appears that these relationships may partially be related to task duration. Longer duration tasks may prompt more lax lanekeeping, either due to workload effects or to the continued effort required for “crisp” vehicle control over longer periods. The relationships may also reflect, in part, some shift of attention from just lanekeeping to something else such as event monitoring (as opposed to just lanekeeping). These relationships require further analysis and study. Median SDLP was negatively correlated with eyeglance metrics related to task-related glances (duration, rate, and percent of task time spent glancing at task-related areas) and glances to the “not road” category. In other words, the higher the percentage of time spent looking at task-related locations during the task and to any area categorized as not road, the lower the median SDLP. These relationships seem counterintuitive, since it is expected that the more glances away from the road, the more variable lane position would become. This was not observed in these correlations. Also, there was a negative correlation between glance rate per second to the road-and-median SDLP (the higher the glance rate to the road, the less the SDLP). Whether this suggests that more frequent visual sampling of lane position, versus steady gazing ahead, is associated with less variability in lane position, or whether it suggests that higher glance rates are associated with steering holds is not known and deserves further study.

Table 4-6. Correlations Between Eyeglance Metrics and Reliable Driving Performance Metrics Across the Full Task Set for the Road

Correlations for All Tasks from Road Data						
	Median SDLP	*Median Speed Diff*	*%Cross Trials*	*%LVD Miss Rate*	*%CHMSL Miss Rate*	*%FVTS Miss Rate*
MeanTskglncs	0.822	0.870	0.559	-0.815	-0.768	-0.735
MeanTaskdur	0.780	0.949	0.625	-0.765	-0.847	-0.663
MeanmeanTdur	0.528	0.802	0.550	-0.577	-0.779	-0.306
MeansdTdur	0.583	0.843	0.600	-0.623	-0.796	-0.330
MeanTglsprs	-0.693	-0.870	-0.652	0.792	0.898	0.505
MeanglncesRD	0.816	0.874	0.560	-0.810	-0.773	-0.732
MeanduratRD	0.764	0.952	0.629	-0.750	-0.855	-0.644
MeanmeanRDdr	0.577	0.837	0.578	-0.621	-0.810	-0.355
MeanmedRDdur	0.520	0.786	0.533	-0.600	-0.788	-0.363
MeansdRDdur	0.620	0.856	0.618	-0.675	-0.808	-0.349
MeangrateRD	-0.717	-0.877	-0.661	0.808	0.903	0.518
MeanpctdurRD	0.707	0.839	0.634	-0.807	-0.894	-0.510
MeanglncesSA	0.738	0.891	0.546	-0.747	-0.851	-0.723
MeanduratSA	0.733	0.883	0.544	-0.741	-0.850	-0.720
MeanmeanSAdr	0.755	0.867	0.662	-0.682	-0.755	-0.517
MeanmedSAdur	0.724	0.839	0.676	-0.649	-0.730	-0.474
MeanpctdurSA	0.672	0.792	0.508	-0.714	-0.878	-0.766
MeanglncesTR	-0.101	-0.552	-0.352	0.188	0.623	0.236
MeanduratTR	-0.243	-0.634	-0.473	0.299	0.698	0.263
MeanmeanTRdr	-0.732	-0.866	-0.847	0.688	0.874	0.345
MeangrateTR	-0.727	-0.935	-0.860	0.746	0.920	0.511
MeanpctdurTR	-0.774	-0.914	-0.862	0.773	0.907	0.485
MeangrateNA	-0.863	-0.789	-0.600	0.911	0.745	0.549
MeanpctdurNA	-0.852	-0.770	-0.566	0.893	0.724	0.528
MeanglncesMR	0.741	0.903	0.547	-0.750	-0.851	-0.720
MeanduratMR	0.739	0.898	0.543	-0.745	-0.851	-0.717
MeanmeanMRdr	0.778	0.865	0.643	-0.699	-0.767	-0.518
MeanmedMRdur	0.743	0.828	0.639	-0.665	-0.746	-0.489
MeanglncesNR	0.827	0.867	0.548	-0.819	-0.763	-0.728
MeanduratNR	0.787	0.792	0.501	-0.786	-0.660	-0.700
MeanmeanNRdr	-0.707	-0.747	-0.557	0.847	0.856	0.502
MeansdNRdur	-0.701	-0.777	-0.569	0.794	0.876	0.510
MeangrateNR	-0.670	-0.865	-0.645	0.775	0.894	0.502
MeanpctdurNR	-0.712	-0.841	-0.636	0.811	0.897	0.523
MaxTdur	0.709	0.902	0.583	-0.684	-0.742	-0.477
MaxRDdur	0.709	0.902	0.583	-0.684	-0.742	-0.477
MaxTRdur	-0.659	-0.771	-0.716	0.597	0.816	0.415

Items in green have + "r" values > 0.707 (original cutoff for repeatability)
Items in yellow have "r" values < - 0.707
Items in blue have + "r" values >0.665 and p<0.05
Items in light yellow have "r" values > - 0.665 and p<0.05

The Speed Difference variable correlated in a similar way with glance variables. It was related very consistently with the mean number of glances to the road and their durations and with the mean number of glances to the mirrors and their durations. (See Table 4-6). The latter is measured in two ways, one using the SA location type, which includes both glances to mirrors and speedometer, and one with the MR location type, which includes glances to mirrors only. Of the two, the correlations with mean glances and durations to the mirrors are slightly stronger in this case. Median Speed Difference is correlated strongly with eye metrics associated with the overall task (e.g., with the variables called Mean Task Glances and Mean Task Duration). The positive correlations with the road and mirror glance metrics may perhaps be interpreted to indicate that drivers were adjusting speed in relation to their situation awareness, as developed from glances to the road and mirrors/speedometer. In these relationships, the more glances and the longer the glances, however, the larger the speed difference during the task. On the face of it, this seems like a somewhat counterintuitive result. However, Speed Difference was also largely driven by task duration. It was constrained, as was SDLP, for shorter tasks. Instead, longer duration tasks may prompt laxer longitudinal control, either due to workload effects or to the continued effort required for “crisp” vehicle control. As suggested previously, it is also possible that these correlations may hint at a state of monitoring for events, rather than a state of monitoring speed or lanekeeping (i.e., driving), but such a hypothesis requires further analysis and study for verification. Three

of the four highest negative correlations involved metrics related to glance rate—glance rate to the road, glance rate to task-related locations, and total glance rate across all locations. These relationships indicated that the higher the glance rate, the lower the speed difference during the task. Given that the relationships to glance rate involve multiple locations, it becomes less likely that the rate of glancing to one location (e.g., the lead vehicle on the road ahead) is benefiting speed maintenance and more likely that during periods of very high glance rates to multiple locations (road, task, and other). There are "holds" on the accelerator pedal for short periods of time (possibly as a means of managing workload).

Surprisingly, for the variable of Percent Cross Trials (percent of trials with a cross of the lane line), only negative correlations proved significant and all of these were with task-related glance variables. This outcome was unexpected and contrary to any hypothesized outcome. The correlations suggested that as task-related glance duration, rate, and percent time viewing task increased, as well as maximum glance duration on task-related areas, the Percent Cross Trials decreased.

The measures of driver responsiveness to events (Percent LVD Miss Rate, Percent CHMSL Miss Rate, and Percent FVTS Miss Rate) in Table 4-6 also correlated in a consistent way with the eyeglance measures, though the correlations were the strongest for Percent CHMSL Miss Rate. Among the highest positive correlations were several with glance rate metrics. These included glance rate per second for the total task (all glances included, regardless of location), with glance rate to the road, with glance rate to task-related areas, and glance rate to not-road areas. These relationships indicated that the higher the glance rate, the higher the miss rate. As seen in the graphs depicting the Task by Location Type interaction, high glance rates to the task are associated with high glance rates to the road (the pattern of looking back and forth between task and road). (Note that these measures are correlated and do not provide independent information. For example, glance rates to all locations include glance rates to the road and glance rates to the tasks.) This may be an instance of a relationship specific to a subset of tasks (visual-manual) dominating the full task set in the computation of overall correlations. A possible explanation is that when there is a high glance rate, there tends to be a high number of transitions between locations, and that during these transitions, events tend to be missed (perhaps because vision is suppressed during each transition).

Note that in Figure 4-6 there are also a cluster of correlations related to task-related glance metrics (duration, rate, percent of task time spent looking at task-related areas, and maximum glance duration to task-related areas). Since the vast majority of task-related glances were made during visual-manual tasks, this suggests further that a subset of tasks had a prominent influence in the overall task set on which these correlations were done. Correlations with rate and percent of time looking at NA areas during the task may largely be due to tasks in which paper stimulus materials were held in front of the eyes, obstructing them from being scored, but also obstructing the driver from seeing events in front of them. Finally, a cluster of positive correlations emerged for metrics related to glances made to the not road category of locations indicating that miss rates increased as these glance metrics increased.

Strong negative correlations emerged for the various metrics associated with numbers and durations of glances to the road and to the mirror/situation awareness locations (mean, percent duration in a location, and accumulated duration across task). These negative correlations indicated that the fewer and shorter the glances are to the road and/or mirror locations, the higher the miss rates are for CHMSLs, FVTS, and LVDs. The underlying cause(s) of fewer and shorter glances to these locations are difficult to

interpret. Whether they are related to task loading or whether they are related to changes due to responses to events, particularly during long tasks, or to multiple effects from different types of tasks, is not clear. One possibility is that glance durations to the mirrors become very short when the glance rate between locations is very high (which may occur, for example, during a visual-manual task during event-monitoring). In such a condition, both elevated glance rate and shortened mirror glances may be associated with elevated miss rates for CHMSLs. A separate effect may be one in which there are fewer glances to the mirror during high workload tasks, which also may be associated with miss rates for CHMSLS. These effects may emerge more clearly and be more separable in the analyses of separate task types below. Also, note that for FVTS Miss Rates, only negative correlations with glance metrics were significant (so only fewer and shorter total glance times to the mirrors predicted increased FVTS miss rates).

Correlations Across Subsets by Task Type

To clarify the interpretation of the correlations done across the entire task set, additional correlations were done on smaller subsets of tasks. Specifically, correlations were separately done on the visual-manual tasks and the auditory-vocal tasks with the mixed-mode tasks and Just Drive task combined into a third category. Though Just Drive is quite different from both task types, and is not a mixed-mode task, it was grouped with them to enable examination of the remaining variance after the variances for the visual-manual and auditory-vocal tasks were computed in this series of analyses.

Visual-Manual Tasks

As shown in Table 4-7, when only visual-manual tasks are included in the correlation, for the metric of Median SDLP, there were strong positive correlations with several metrics related to number of glances, including Mean Task glances (all glances to all locations), number of glances to the Road, and number of glances to Situation Awareness areas, mirrors and also not road areas. Also, there was a positive correlation with Mean Task Duration (based on glance information). These relationships indicated that as number of glances increased, and hence task duration increased, so did Median SDLP. Consistent with this, there were also positive correlations with total glance time metrics for Road, Situation Awareness areas, and Mirrors. Negative correlations emerged for task-related glance durations, suggesting that the longer the duration of task-related glances, the lower SDLP. Also there were negative correlations for glance rate and percent of task time spent looking at the NA (not able to be scored) area (often an indication that paper stimulus materials were held in front of the face), and time associated with this area was also associated with lower SDLP.

For the metric of Median Speed Diff, there were positive correlations for Mean Task Glances, number of glances and total glance time to the road, situation awareness areas, and mirrors, and number of task-related glances and number of glances to the not road area. The fact that as number of task-related glances increased, so did speed difference, was rather interesting.

Some strong positive relationships emerged between glance measures and percent trials with a cross of the lane line that were not present for the overall data set. This suggests that the relationships between glance measures and lane departure metrics (often reported previously in the U.S. literature) may be strongest for the visual-manual subset of tasks. The metric of Percent Cross Trials was related to mean number of glances to the road and total glance time to the Road and situation awareness areas, and mean number of glances to task-related areas. It was also positively related to Mean Task Glances (all glances to all locations during task) and Mean Task Duration (based on eyeglanced information).

(The only negative correlations were again with glances associated with the NA area, indicating some obscuration of the eyes, often by paper materials held by the driver between the eyes and camera. This may again indicate that when the driver is looking at the paper materials (and perhaps blocking his or her own view of the road), steering becomes conservative, since the trials with a crossing of a lane line decreased, just as SDLP decreased).

Surprisingly, there are many fewer relationships between eye behavior and responsiveness to events in the data for visual-manual tasks only. However, even though only a few correlations remain, those that were significant were meaningful.

For Percent Lead Vehicle Deceleration Miss Rate, the negative correlation with Task Duration (based on eyeglance data) indicates that the longer the visual-manual task, the lower the miss rate, which is consistent with the fact that some of the visual-manual tasks were so short that the LVD events were not even detectable within the duration of the task. The negative correlation with total glance time to the road indicates that the less time spent viewing the road during a visual-manual task, the higher the LVD miss rate. Along this line, the correlations with number of glances to situation awareness areas, and mirrors (and total glance time to mirrors) indicates that as these decreased, the LVD miss rate increased.

For the Percent CHMSL Miss Rate, there was a high positive correlation with glance rate to task-related areas, indicating that the more glances per second to the task, the higher the miss rate.

For the Percent FVTS Miss Rate, there was a significant negative correlation with number of glances to mirrors, indicating that as the number of mirror glances went down, the miss rate for FVTS events increased (and FVTS events were detected in the mirror).

Table 4-7. Correlations Between Eyeglance Metrics and Other Driving Performance Metrics for Visual-Manual Tasks Only

Correlations: Visual-Manual Tasks Only From The Road						
	Median SDLP	*Median Speed Diff*	*%Cross Trials*	*%LVD Miss Rate*	*%CHMSL Miss Rate*	*%FVTS Miss Rate*
MeanTskglncs	0.721	0.815	0.793	-0.584	-0.056	-0.070
MeanTaskdur	0.894	0.949	0.942	-0.747	-0.220	-0.127
MeanmeanTdur	0.401	0.287	0.321	-0.405	-0.483	-0.136
MeansdTdur	0.518	0.415	0.458	-0.468	-0.492	-0.122
MeanTglsprs	-0.562	-0.446	-0.439	0.582	0.608	0.296
MeanglncesRD	0.706	0.803	0.782	-0.569	-0.039	-0.059
MeanduratRD	0.960	0.948	0.966	-0.809	-0.419	-0.183
MeanmeanRDdr	0.492	0.379	0.397	-0.491	-0.558	-0.220
MeanmedRDdur	0.336	0.210	0.204	-0.404	-0.579	-0.279
MeansdRDdur	0.619	0.516	0.547	-0.552	-0.535	-0.177
MeangrateRD	-0.639	-0.531	-0.514	0.658	0.636	0.335
MeanpctdurRD	0.560	0.444	0.454	-0.539	-0.596	-0.269
MeanglncesSA	0.762	0.776	0.646	-0.674	-0.588	-0.628
MeanduratSA	0.807	0.810	0.721	-0.640	-0.471	-0.492
MeanmeanSAdr	0.226	0.186	0.351	0.085	0.445	0.488
MeanmedSAdur	0.161	0.114	0.296	0.143	0.467	0.509
MeanpctdurSA	-0.036	-0.157	-0.276	0.046	-0.459	-0.575
MeanglncesTR	0.628	0.741	0.754	-0.463	0.089	0.091
MeanduratTR	0.444	0.574	0.556	-0.344	0.177	0.071
MeanmeanTRdr	-0.700	-0.621	-0.661	0.610	0.531	0.171
MeangrateTR	-0.435	-0.290	-0.208	0.530	0.756	0.575
MeanpctdurTR	-0.555	-0.425	-0.405	0.560	0.658	0.376
MeangrateNA	-0.790	-0.786	-0.728	0.878	0.332	0.247
MeanpctdurNA	-0.802	-0.822	-0.786	0.865	0.323	0.235
MeanglncesMR	0.714	0.726	0.562	-0.710	-0.655	-0.730
MeanduratMR	0.806	0.801	0.678	-0.709	-0.559	-0.603
MeanmeanMRdr	0.372	0.314	0.466	-0.038	0.271	0.353
MeanmedMRdur	0.297	0.232	0.399	0.048	0.291	0.361
MeanglncesNR	0.738	0.830	0.809	-0.597	-0.081	-0.078
MeanduratNR	0.525	0.645	0.611	-0.431	0.084	-0.004
MeanmeanNRdr	-0.698	-0.592	-0.585	0.660	0.660	0.381
MeansdNRdur	-0.697	-0.611	-0.648	0.567	0.581	0.302
MeangrateNR	-0.488	-0.365	-0.365	0.512	0.576	0.269
MeanpctdurNR	-0.578	-0.460	-0.465	0.556	0.614	0.304
MaxTdur	0.425	0.347	0.479	-0.254	-0.100	0.037
MaxRDdur	0.425	0.347	0.479	-0.254	-0.100	0.037
MaxTRdur	-0.401	-0.253	-0.293	0.252	0.474	0.258
			Items in green have + "r" values > 0.707 (original cutoff for repeatability)			
			Items in yellow have "r" values < - 0.707			
			Items in blue have + "r" values >0.665 and p<0.05			
			Items in light yellow have "r" values > - 0.665 and p<0.05			

Auditory-Vocal Tasks

Correlations for the auditory-vocal tasks (Table 4-8), surprisingly, showed similar relationships to Median SDLP. This provides further indication that the relationship is not dependent upon task-related glances being made, since there are so few task-related glances made during auditory-vocal tasks. In fact, the relationship between number of glances and total glance time to the road and mirrors/situation awareness areas are as strong or stronger in the auditory-vocal subset than in the full set of tasks or in the visual-manual subset. These categories of metrics appear to be contributing most heavily to the overall relationship of Mean Task Glances (for the whole task) to SDLP. The number of glances can increase as task duration increases, and indeed there is again a strong correlation between Mean Task Duration (based on eye data) and Median SDLP. These relationships suggest that the more glances to the road and mirrors, and the longer the task is, the larger SDLP is. Though it makes sense that SDLP grows with task duration (as discussed in Chapter 3), the fact that more gazing at the road results in larger SDLP seems somewhat counterintuitive. It may be suggestive of an underlying “satisficing” process in which the driver feels more aware of the road and relaxes lanekeeping tolerances somewhat, especially as the task lengthens (auditory-vocal tasks, with one exception, were ~2 minutes in length). Alternatively, longer duration tasks may provide less driver discretion and so may be associated with laxer vehicle control because of workload effects (for auditory-vocal tasks) “lost in thought” distraction effects (for Just

Drive), or the continued effort required for "crisp" vehicle control. This is unclear and deserves further investigation.

Of interest, however, was what happened with a metric that expressed glance time at the road as a percent of task time spent looking at the road. It produced a negative correlation with Median SDLP (as well as Speed Diff). The fact that the direction of the correlation changed when total glance time to the road was divided by task duration and was thus expressed as a proportion or percent is very interesting. What it indicates is that when expressed this way, with the time-component of the metric divided out as the percent of task time looking at the road increased, SDLP and Speed Diff decreased. This would seem to make more sense that as a greater percentage of the task is spent with eyes-on-the road, there is less deviation in lane position and less difference in speed. This result for auditory-vocal tasks stood in striking contrast to the correlation for the full set of tasks, which yielded positive rather than negative correlations between MeanpctdurRD and Median SDLP as well as Median Speed Diff.

There were no strong meaningful correlations with Percent Cross Trials in the data for auditory-vocal tasks.

There were limited correlations between the event miss variables and eyeglance measures for the auditory-vocal tasks, but those that were significant were meaningful. Although there were some hints of correlations with task-related glance metrics, these correlations are based on so few observations (many fewer observations than any other cells in the matrix, since only some auditory-vocal tasks led to upward glances, and even those led to only one or two glances on average). Therefore, it is not known whether these correlations are stable or meaningful.

For Percent LVD Miss Rate, two significant positive correlations emerged, both related to glance duration. These were Mean Tdur (or the mean glance duration across glances of all types to all locations during the task) and Mean Road Dur (or mean duration of glances to the road). The longer these were, the higher the LVD Miss Rate was (though it is important to keep in mind that the miss rate for LVD events was lower for auditory-vocal tasks than for visual-manual tasks). Nonetheless, the finding is consistent with other findings on auditory-vocal tasks, especially given that their variance was examined separately from the other tasks in this analysis.

For Percent CHMSL Miss Rate, significant negative correlations emerged with glance durations to situation awareness areas and mirrors, such that as glance duration on the mirrors decreased, CHMSL Miss Rate increased. This is puzzling and it is not clear how to interpret this. There was a negative correlation as well with glance duration to not road areas, suggesting that the longer the duration glances to not road areas, the higher the miss rate. This metric may be an indirect indicator of a more active scanning state during some auditory-vocal tasks than others that may be associated with steady gazing. Finally, there were negative correlations with the NA area (glance rate and percent of task time spent looking). It is not clear whether these NA correlations are meaningful.

For Percent FVTS Miss Rate, there was a significant positive correlation with percent of task duration spent looking at the Road (MeanpctdurRD). This is very interesting, since when this measure is high, there tended to be reduced scanning of the mirrors. Indeed, consistent with this, there were numerous negative correlations relating number of glances to mirrors and total glance time to mirrors to Percent FVTS Miss Rate.

Table 4-8. Correlations Between Eyeglance Metrics and Driving Performance Metrics for Auditory-Vocal Tasks Only

Correlations: Auditory-Vocal Tasks Only from Road Data						
	Median SDLP	*Median Speed Diff*	*%Cross Trials*	*%LVD Miss Rate*	*%CHMSL Miss Rate*	*%FVTS Miss Rate*
MeanTskglncs	0.918	0.833	0.143	-0.476	-0.552	-0.855
MeanTaskdur	0.928	0.882	0.206	-0.184	-0.617	-0.716
MeanmeanTdur	-0.219	-0.092	-0.062	0.733	-0.178	0.601
MeansdTdur	-0.097	0.054	0.014	0.608	-0.125	0.613
MeanTglsprs	0.479	0.324	-0.013	-0.578	-0.124	-0.845
MeanglncesRD	0.917	0.832	0.143	-0.479	-0.551	-0.853
MeanduratRD	0.919	0.880	0.211	-0.086	-0.619	-0.684
MeanmeanRDdr	-0.113	0.011	-0.032	0.695	-0.215	0.537
MeanmedRDdur	-0.281	-0.255	-0.114	0.589	-0.108	0.403
MeansdRDdur	-0.102	0.049	-0.011	0.420	0.076	0.664
MeangrateRD	0.339	0.181	-0.072	-0.585	-0.012	-0.766
MeanpctdurRD	-0.828	-0.706	-0.200	0.529	0.452	0.947
MeanglncesSA	0.919	0.835	0.127	-0.423	-0.561	-0.852
MeanduratSA	0.937	0.857	0.157	-0.392	-0.589	-0.851
MeanmeanSAdr	0.923	0.907	0.381	0.387	-0.729	-0.688
MeanmedSAdur	0.850	0.844	0.504	0.474	-0.756	-0.607
MeanpctdurSA	0.858	0.744	0.173	-0.474	-0.497	-0.945
MeanglncesTR	-0.251	-0.678	0.064	-0.609	0.404	-0.981
MeanduratTR	0.046	-0.431	-0.233	-0.816	0.656	-0.995
MeanmeanTRdr	0.817	0.448	-0.911	-0.957	0.999	-0.528
MeangrateTR	-0.280	-0.699	0.094	-0.585	0.377	-0.975
MeanpctdurTR	0.042	-0.434	-0.229	-0.814	0.653	-0.995
MeangrateNA	-0.867	-0.852	-0.132	-0.528	0.739	0.621
MeanpctdurNA	-0.864	-0.855	-0.038	-0.550	0.709	0.581
MeanglncesMR	0.915	0.835	0.097	-0.430	-0.542	-0.846
MeanduratMR	0.938	0.864	0.115	-0.402	-0.564	-0.840
MeanmeanMRdr	0.937	0.944	0.308	0.318	-0.684	-0.638
MeanmedMRdur	0.898	0.902	0.422	0.137	-0.626	-0.580
MeanglncesNR	0.924	0.849	0.121	-0.471	-0.542	-0.828
MeanduratNR	0.927	0.858	0.164	-0.359	-0.607	-0.834
MeanmeanNRdr	0.839	0.824	0.349	0.547	-0.859	-0.716
MeansdNRdur	0.733	0.741	0.070	0.036	-0.576	-0.644
MeangrateNR	0.642	0.517	-0.034	-0.664	-0.138	-0.837
MeanpctdurNR	0.856	0.761	0.186	-0.552	-0.465	-0.902
MaxTdur	0.610	0.598	-0.079	-0.102	-0.173	-0.239
MaxRDdur	0.610	0.598	-0.079	-0.102	-0.173	-0.239
MaxTRdur	0.968	0.735	-0.998	-0.793	0.915	-0.194

Task-related glances were so few (<1 or 2 per task) and restricted to only some tasks, that meaning of these correlations is questionable.

Items in green have + "r" values > 0.707 (original cutoff for repeatability)
Items in yellow have "r" values < - 0.707
Items in blue have + "r" values >0.665 and p<0.05
Items in light yellow have "r" values > - 0.665 and p<0.05

Mixed-Mode Tasks With Just Drive

With only a single mixed-mode task tested on the road, it was not possible to examine a subset of mixed-mode tasks for correlations with the road data. However, it can be inferred from the patterns in Table 4-7 and Figure 4-8, that the Just Drive task also has a great deal of influence on the correlations between eyeglance measures in the overall data set and the event-miss measures.

Insofar as the Just Drive task involved driving and monitoring for events, it may be that drivers placed more emphasis or allocated more attention to CHMSL, FVTS, and LVD event monitoring than in other conditions. If that were the case, then it might be expected that the relationships between eyeglance measures and event detection performance would be strongly influenced by the Just Drive tasks.

To review, the findings from correlational analyses between eyeglance metrics and other driving performance metrics (of lanekeeping, speedkeeping, and event detection) can be summarized in terms of several major clusters of effects:

- Metrics associated with glances to the road and situation awareness areas (mirrors and speedometer) tended to be related to lanekeeping (Median SDLP) and speedkeeping (Median Speed Difference). The nature of these relationships, however, were surprising. It was not the case that the more glances to the road, the better the lanekeeping (or the smaller the SDLP). Rather, in general, the more glances to the road, the greater the standard deviation in lane position and the greater the speed difference. It appears that this may partially be related to time-based processes, but it may also in part reflect some shift of attention from "just lanekeeping and speedkeeping" to something else. There may have been a shift of attention to active monitoring of the roadway, perhaps for event detection (in addition to performance of in-vehicle tasks). More glances to mirrors were related to increased speed difference. The one exception was a negative relationship between the time-independent metric of Percent of Task Time Spent Looking at Road for auditory-vocal tasks and Median SDLP as well as Median Speed Difference, such that higher percent time looking at road led to decreased SDLP and Speed Difference.
- Metrics associated with glances to task-related areas were related to miss rates for events (such as mean glance rate to task-related areas, and percent of task duration spent viewing task-related areas). Mean Number of Glances to Task Related Areas was also associated with excursions from the lane (Percent of Trials with a Cross of the Lane Line).
- Metrics associated with glances to mirrors and situation awareness areas were associated with more missed CHMSL and FVTS events (the latter of which appeared in the left outside mirror). For example, total glance time to mirrors and duration of glances on mirrors, was correlated with missed CHMSLs and FVTS events. However, the correlations were negative (e.g., for FVTS events, as total glance time on the mirrors went down, the miss rate went up).

4.4.4 Summary of Findings from On-Road Eyeglance Data

The key findings of the eyeglance data analyses are:

3. Several categories of eyeglance measures proved reliable in split-half analyses of the data:
    - Number of glances
        - To road locations
        - To situation awareness locations
        - To task-related areas
    - Durations of glances
        - Mean (for most locations – road, situation awareness, task (borderline), and not road)
        - Median (for some locations – road, situation awareness)
        - Standard deviation (for some locations – road, task, total/all, not road)
        - Max (for only certain location types – road, task, total/all)

- Accumulations of durations for certain location types
  - Total glance time to road location
  - Total glance time to situation awareness location
  - Total glance time to task-related areas
- Percents (or proportions) of task time spent looking at a location type
  - To road locations
  - To situation awareness areas (borderline)
  - To task-related areas
- Rates of glances per second
  - Overall (total/all), road, task, not road)

As with the track findings, these same measures tended to reveal interesting findings in the road eye data. First, and very important among these findings, was the fact that not all information is in the simple classification of glances as on-road or off-road. Glances to road, task, and mirror locations all carried important information. Among these, there were measures that discriminated between types of tasks. There were distinct patterns of glancing revealed across types of locations (road, mirrors, task). In the road data, as in the track data, among the most interesting findings from formal statistical analysis was a significant Task by Location Type interaction across many of the eyeglance measures. Notable was the fact that the pattern of glances to the roadway discriminated task types particularly well, and a measure that integrated multiple measures together—Proportion of Task Time Spent Looking at the Road (Pct Dur Rd)—was particularly useful for characterizing patterns of glancing associated with tasks, along with a similar measure applied to each other glance location (task and mirrors).

In summary, the effects observed on the road for in-vehicle tasks on eyeglance behavior replicated those observed on the test track. Eyeglance metrics showed distinct patterns for different types of task engagement (just driving versus concurrently performing an auditory-vocal task or concurrently performing a visual-manual task). The Just-Drive task was distinguished by patterns in which drivers looked at the road about 81 percent of the time and scanned their mirrors about 15.4 percent of the time. Glances on the road were about 3.8 seconds in duration, on average. Auditory-vocal tasks showed a somewhat similar pattern, though drivers gazed at the forward roadway somewhat more (87%), using longer gazes (4 to 8 seconds, on average), and scanned their mirror somewhat less (11.4%). This miss rate for event detection was slightly elevated over just driving for auditory-vocal tasks for CHMSL and LVD events (showing an increase of ~6% for CHMSL and ~4% for LVD events), and somewhat more for the peripheral FVTS (an increase of ~14%), though event detection was less affected by auditory-vocal tasks than by visual-manual tasks. Visual-manual tasks showed a different pattern, in which drivers looked at the forward roadway much less (viewing the road only 42% to 68% of the time during a task), and using glance durations on the road that were less than 2 seconds long, on average. This reduction in glances to the road was made in order to view task-related areas required for performing the in-vehicle activity (viewing the task 24% to 52% of the time during its length). For visual-manual tasks, glances tended to cycle frequently back-and-forth between the task and the roadway locations, and glance-rate measures proved to carry interesting information. Visual-manual tasks led to a more pronounced reduction in mirror-scanning (to 6.4%) and were associated with higher rates of missed events (though this was sometimes due to a methodological constraint for

LVDs). Increases in miss rates over Just Drive were approximately 23 percent for CHMSLs, 28 percent for LVDs, and 65 percent for FVTS events on average.

1. Interrelationships with driving performance measures revealed:
   - Correlations with SDLP
   - Correlations with Speed Diff
   - Correlations with Event Detection (due to influence of Just Drive and selected tasks)
2. A striking new finding emerged from relating eyeglance data to event-detection data. Qualitative exploration of the time series data suggested that event-detection affected eyeglance behavior. In brief, formal analyses confirmed that when an event occurred and was responded to, eyeglance behavior changed. However, the patterns of change in the road data were different from those in the track data.
   - For CHMSL events:
     - Durations of glances increased but not for situation awareness locations
     - Rate of glancing decreased slightly to road and task-related areas, but increased to situation awareness areas
   - For LVD events:
     - Durations of glances to the road decreased
     - Rate of glancing decreased to the Road and Task-Related but increased to Situation Awareness areas
   - For FVTS events:
     - Durations of glances decreased
     - Rate of glancing to road and task-related areas decreased, but increased to situation awareness areas
3. Changes to glance durations interacted with task type and were more pronounced for Just Drive and auditory-vocal tasks than for visual-manual tasks, which usually showed a different pattern. Events, when detected, appeared to act as attentional interrupts for auditory-vocal tasks and the Just-Drive task, in eliciting more active scanning of the forward roadway and mirrors (a strategy that would be expected to improve subsequent event detection). Event detection also affected glance behavior during visual–manual tasks, but somewhat differently: rate of scanning between all locations (road, task, and mirrors) increased, but higher glance rates were associated with higher rates of missed events (except for LVD events).
4. The finding that event detection affects glance behavior has implications for analysis and design of future studies. Methods used to study event detection may affect the behavior of interest and suggest that when evaluating the visual demand of tasks in an advanced information system or in-vehicle device, it is important that multiple test trials be conducted—some with and some without event detection. The trials used to evaluate the visual demand of a task should NOT include events to-be-detected in order to obtain "clean" measurements of glance behavior, free from the influence of co-occurring events.

   The findings that event detection may affect glance behavior has implications for future research.

5. Recommendations on eyeglance metrics for use in future work. The usefulness of the traditional eyeglance metrics for visual-manual tasks was confirmed through analyses of repeatability, predictive validity, and discriminability (described in a later section of this chapter). These included: "number of glances to task-related areas" and "total glance time to task." Retention of "duration of task-related glances" was recommended as well. Additional metrics (which emerged from new findings from this research) were recommended for use in future research on visual-manual tasks. No eyeglance metrics were recommended for application to the assessment of auditory-vocal tasks, though for research purposes, metrics emerging from this work as promising were identified (such as Proportion of Task Duration Spent Looking at the Road, and Mean Duration of Glances to The Road, and Proportion of Task Duration Spent Looking At Mirrors/Situation Awareness areas).
6. Eyeglance behavior appears to be a key diagnostic for workload, and its associated metrics offer promise as key discriminators in identifying tasks that interfere with visual performance on the road.

4.5 Road Task Effects on Lateral Control

4.5.1 Standard Deviation of Lane Position (SDLP)

Figure 4-41 presents the median SDLP for the 16 DWM tasks evaluated on the road. The range of median SDLP values was small, between about 0.45 ft and 0.65 ft. Higher SDLP values are associated with the auditory-vocal tasks and Just Drive tasks, i.e., the longer duration tasks. There were nonetheless exceptions. SDLP for the Manual Dial task and the Voice-Dial task were approximately equal to Just Drive. Several auditory-vocal tasks had median SDLP values lower than Just Drive. The Book-on-Tape Summarize task had a median SDLP value within the range of the visual-manual tasks.

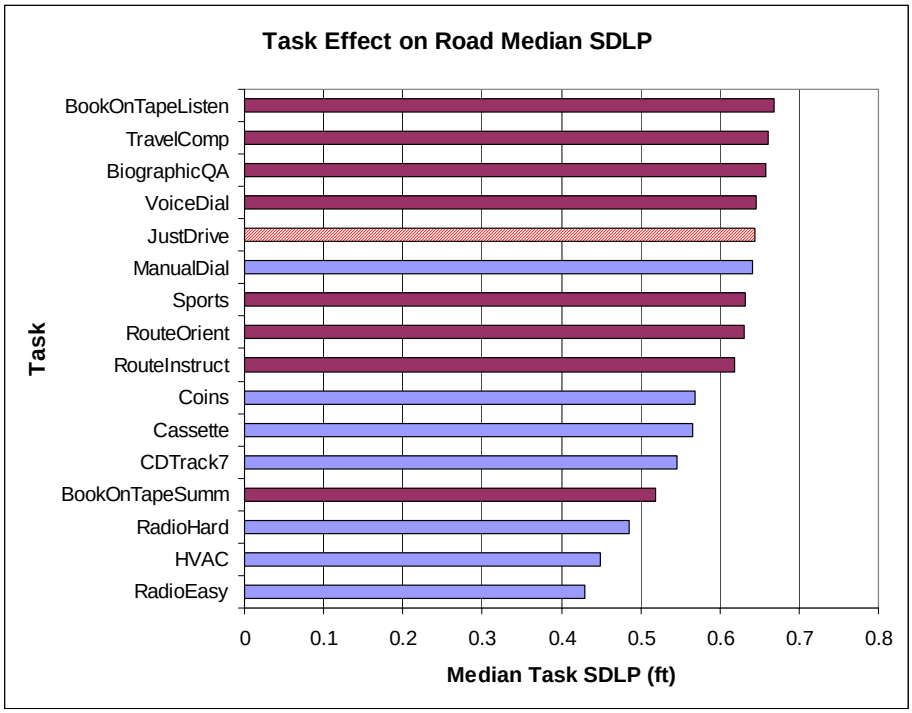


Figure 4-41. Road Median Standard Deviation of Lane Position by Task

Task duration may shed additional light on these results. Figure 4-42 shows the median task durations for the DWM road tasks. Tasks were clearly segregated by type on the road. The visual-manual tasks were the shorter tasks, the auditory-vocal tasks were the longer tasks, and the Book-on-Tape Summarize and the mixed-mode Voice Dial tasks were in between.

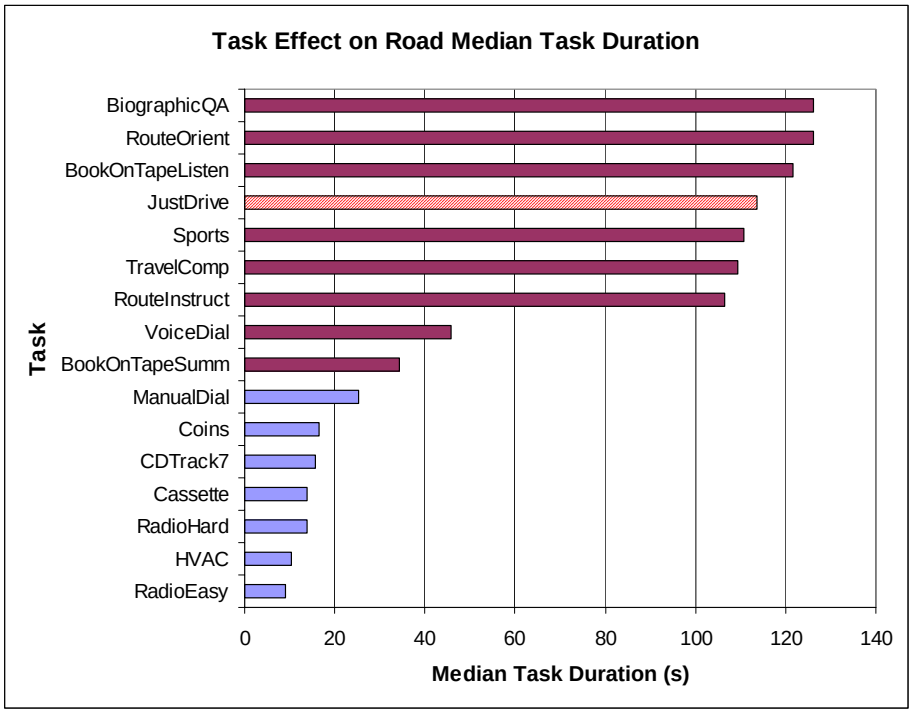


Figure 4-42. Road Median Task Duration by Task

Figure 4-43 presents a plot of road median SDLP as a function of road median task duration for all tasks. Each data point has been labeled with the task name for reference. The plot shows the longer auditory-vocal and Just Drive tasks on the far right. These tasks had median task durations of between 107 seconds and 126 seconds. These tasks had median SDLPs from 0.62 to 0.67 ft. Moving left, the Voice Dial task was found to have a median SDLP of 0.65 ft, within the range of the longer auditory-vocal and Just Drive tasks. Yet the Voice Dial task had a median task duration of approximately 46 seconds, less than half that of the longer task durations. This suggests that longer tasks can have typical SDLP values similar to those of a task half as long. This held for DWM task durations within the 45 seconds to 125 seconds range of durations. In general, longer duration tasks need not necessarily be associated with higher SDLP values.

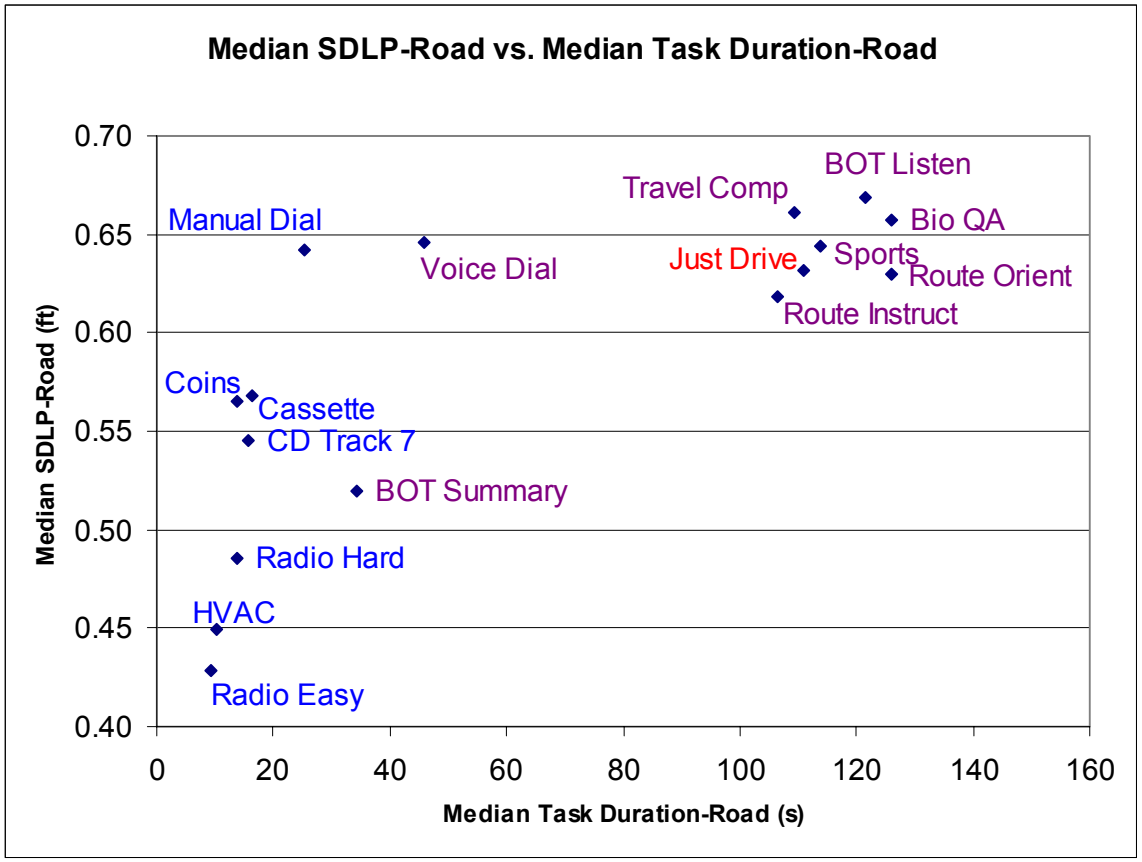


Figure 4-43. Road Median SDLP Versus Road Median Task Duration for All Tasks Performed on the Road

Consider next the Manual Dial task. Manual Dial had a median SDLP comparable to that of Voice Dial (0.64 ft). Yet, the Manual Dial task had a median task duration of 25 seconds, almost half the typical duration of Voice Dial and almost one-fourth the duration of the longer auditory-vocal tasks. This indicates that shorter duration DWM tasks could have typical SDLP values similar to those of tasks between roughly 2 to 4 times greater. This held for task durations between the 25-second and 125-second range. Shorter duration tasks need not necessarily be associated with smaller SDLP values.

Consider the Book-on-Tape Summarize task, with a median task duration of about 35 seconds. Its associated median SDLP was about 0.5 ft. This is about 20 percent less than that of Manual Dial, despite the Book-on-Tape Summarize task being longer by about 40 percent. This suggests that longer tasks need not have even as great an SDLP as a shorter task within the 25-second to 35-second range.

A final point is in regard to Figure 4-44, which presents a more detailed view of the visual-manual tasks alone. Figure 4-44 shows a plot of median SDLP versus Median Task Duration for visual-manual tasks only. A linear trend is evident. As median task durations increased for these tasks, median SDLP values increased as well. However, the Insert Cassette task was an exception. Its typical SDLP was higher than Radio (Hard), Coins, and CD/Track 7 of comparable duration

(about 14 to 16 seconds). The cassette player was positioned lower in the center stack than the radio and this may have contributed to the effect. Nonetheless, this result suggests that median SDLP values can be different even for roughly the same task duration values, at least with tasks in the same class.

These various observations suggest that SDLP can be a useful measure of task load. It is not fully redundant with task duration. There may be different SDLP values for tasks with roughly the same task duration. There can be the same SDLP values for tasks of different task duration. Tasks of different duration can have different and monotonically increasing SDLP values. Finally, tasks of roughly the same task duration can have roughly the same SDLP values.

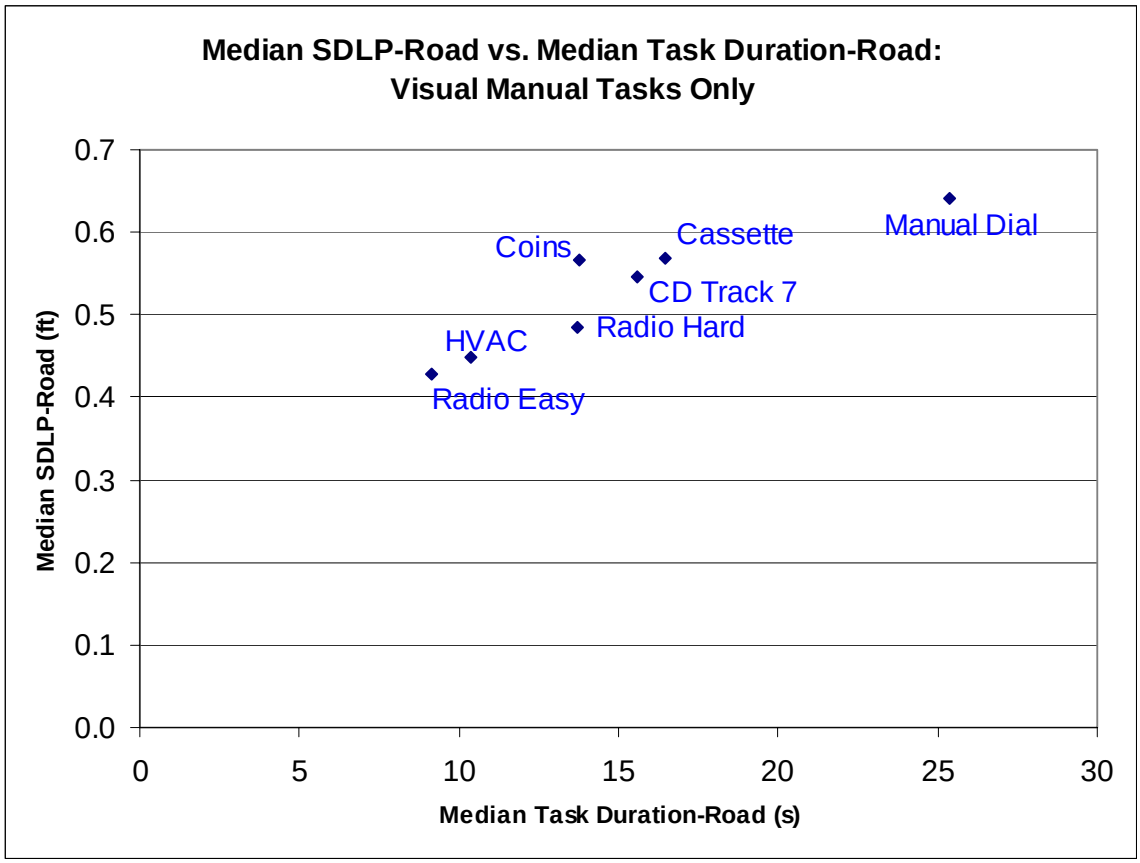


Figure 4-44. Road Median SDLP as a Function of Median Task Time for Visual-Manual Tasks Only

Table 4-9 contains the repeatability results for the selected road measures, including SDLP. As indicated, the repeatability correlation is very good for this measure. Figure 4-45 provides a more detailed view of the repeatability of SDLP. The visual-manual tasks are ordinally related except for Coins. On the other hand, the auditory-vocal tasks are clumped at the high end of the graph. This suggests that these tasks' median SDLP values are randomly varying.

Table 4-9. Repeatability of Selected Road Measures

Driving Measure	Split Group Level Correlation, r	Split Group R^2 %	Estimated Stdev about regression line, S	P-Value, Sig Value
Mean Task Duration	0.999	99.7	2.574	0.000
Median Task Duration	0.999	99.9	1.847	0.000
Mean SDLP	0.891	79.4	0.040	0.000
Median SDLP	0.915	83.7	0.040	0.000
Mean Speed Diff	0.973	94.6	0.654	0.000
Median Speed Diff	0.984	96.9	0.497	0.000
Pct Cross Trials	0.762	58.1	1.072	0.001
Mean Cross Duration	-0.532	28.4	1.113	0.113
Median Cross Duration	0.109	1.2	1.910	0.764
Pct LVD Miss Rate	0.665	44.2	6.865	0.007
Mean LVD RT	0.139	1.9	0.453	0.621
Median LVD RT	0.089	0.8	0.440	0.753
Pct CHMSL Miss Rate	0.858	73.6	5.584	0.000
Mean CHMSL RT	0.008	0.0	0.195	0.976
Median CHMSL RT	0.675	45.6	0.123	0.004
Pct FVTS Miss Rate	0.913	83.3	5.897	0.000
Mean FVTS RT	0.146	2.1	0.684	0.589
Median FVTS RT	-0.235	5.5	0.363	0.381

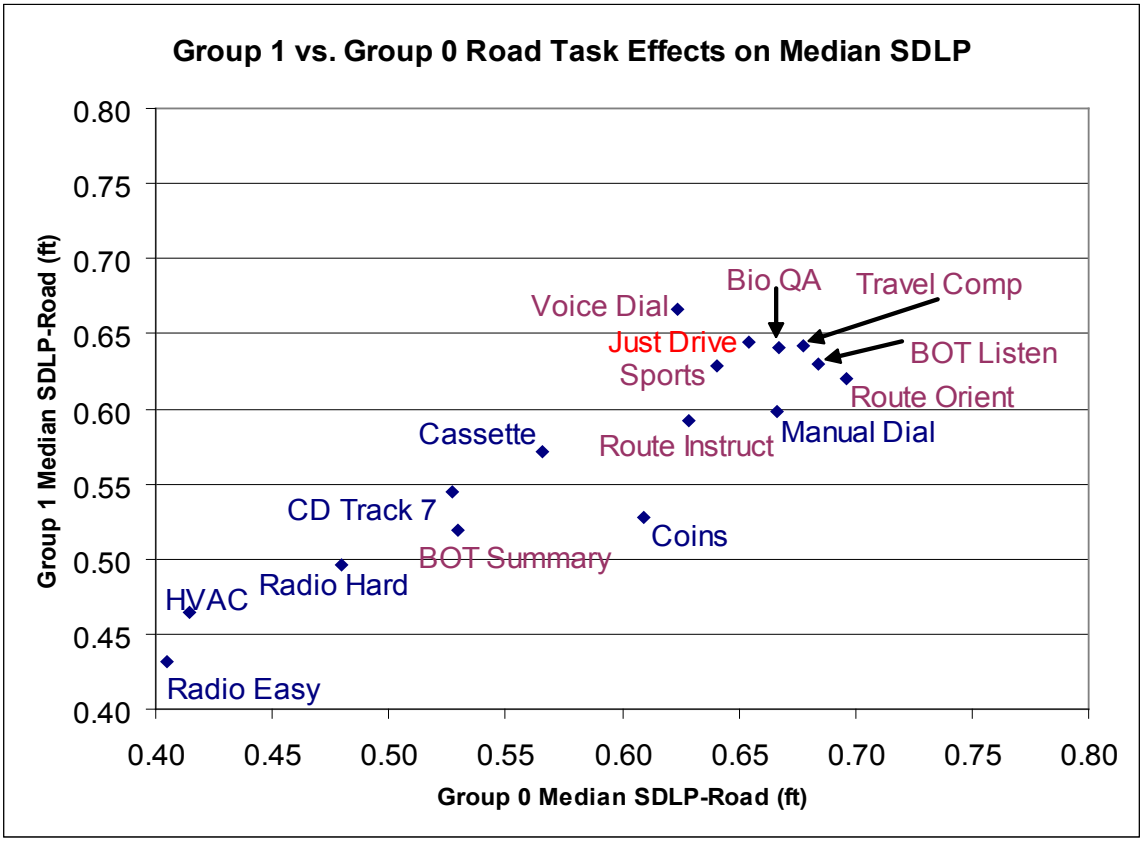


Figure 4-45. Plot of Group 0 Versus Group 1 Road Median SDLP

The next question is how sensible the SDLP outcomes were with respect to prior predictions. A discriminability test was conducted with the Sign Test to determine if SDLP values tended to be larger with the Manual Dial task (a higher-workload task) than with the remaining six visual-manual tasks (lower-workload tasks per prior prediction). Of the six paired comparisons, four were statistically significant ($p \leq 0.05$). The Insert Cassette and CD/Track 7 tasks were not statistically significantly different from Manual Dial in terms of SDLP. However, they were associated with smaller median SDLP values than Manual Dial task, in accord with prior prediction. With six paired comparisons, all six were directionally correct with respect to prior prediction and four were statistically significantly different, despite the small range of SDLP values.

A similar discriminability analysis was carried out with the auditory-vocal tasks. In terms of prior prediction, only three out of 15 paired comparisons were statistically significant in terms of higher SDLP values for higher-workload tasks. All three of these paired comparisons were between the higher-workload auditory-vocal tasks (Route Instructions, Route Orientation, and Travel Computations) versus the Book-on-Tape Summarize task. Of the remaining 12 paired comparisons, nine were not directionally correct (See Figure 4-46). This pattern of results suggests that the SDLP measure may be useful to distinguish among visual-manual tasks but is less so for the auditory-vocal and Just Drive tasks.

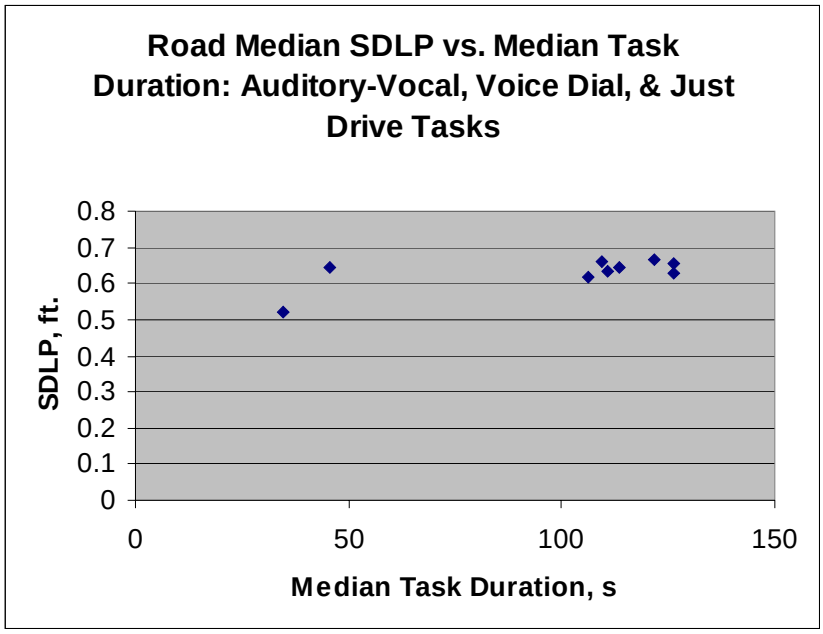


Figure 4-46. Road Median SDLP as a Function of Median Task Duration for Auditory-Vocal Tasks

4.5.2 Percent Lane Exceedance (Cross) Trials

Figure 4-47 shows the ranking of tasks in terms of Percent Lanex (Cross) cases. This is a measure of the percentage of participants with one or more trials that had at least one lane exceedance. This figure shows Sports Broadcast to be the task with the highest incidence of lane exceedance trials. This is not consistent with prior expectations based on the nature of the Sports Broadcast task.

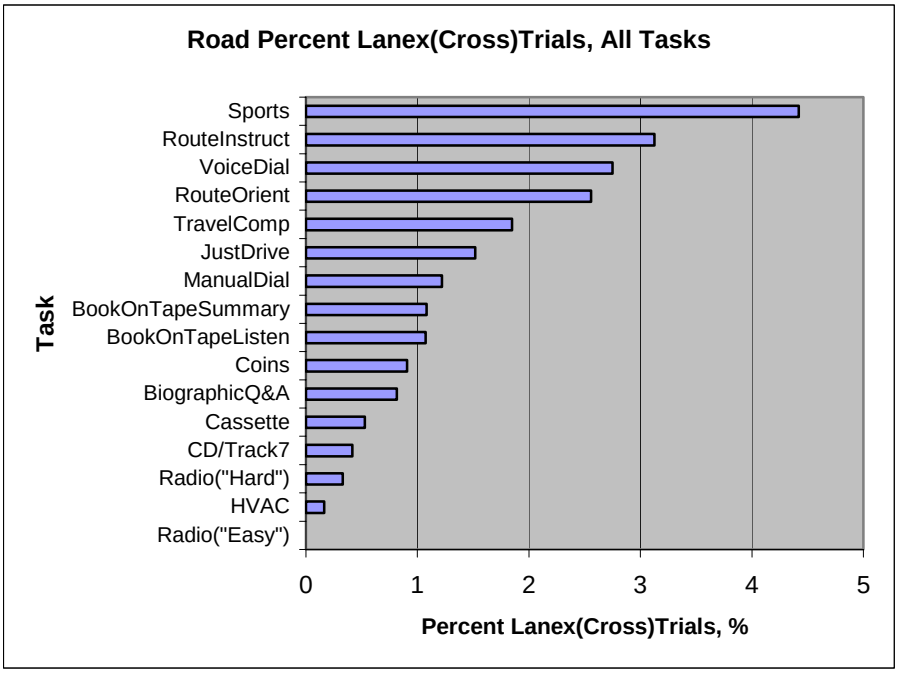


Figure 4-47. Road Percent Lanex (Cross) Trials by Task

Figure 4-48 is a plot of Percent Lanex (Cross) trials as a function of Median Task Duration for visual-manual tasks. Unlike the track data, there does appear to be an increase in lane exceedances as typical visual-manual task durations increase. Lane exceeds were infrequent and the range of Lanex (Cross) percentages is quite small, and smaller for road data compared to track data.

Figure 4-49 is a plot of Percent Lanex (Cross) trials as a function of Median Task Duration for auditory-vocal tasks plus Just Drive. The results are similar to those for the same tasks on the track. There is no systematic relationship between Task Duration and lane exceedance occurrences. There also does not appear to be any systematic effect of the auditory-vocal tasks themselves on this measure of lanekeeping performance. This is consistent with the theory of general versus selective withdrawal of attention mentioned previously.

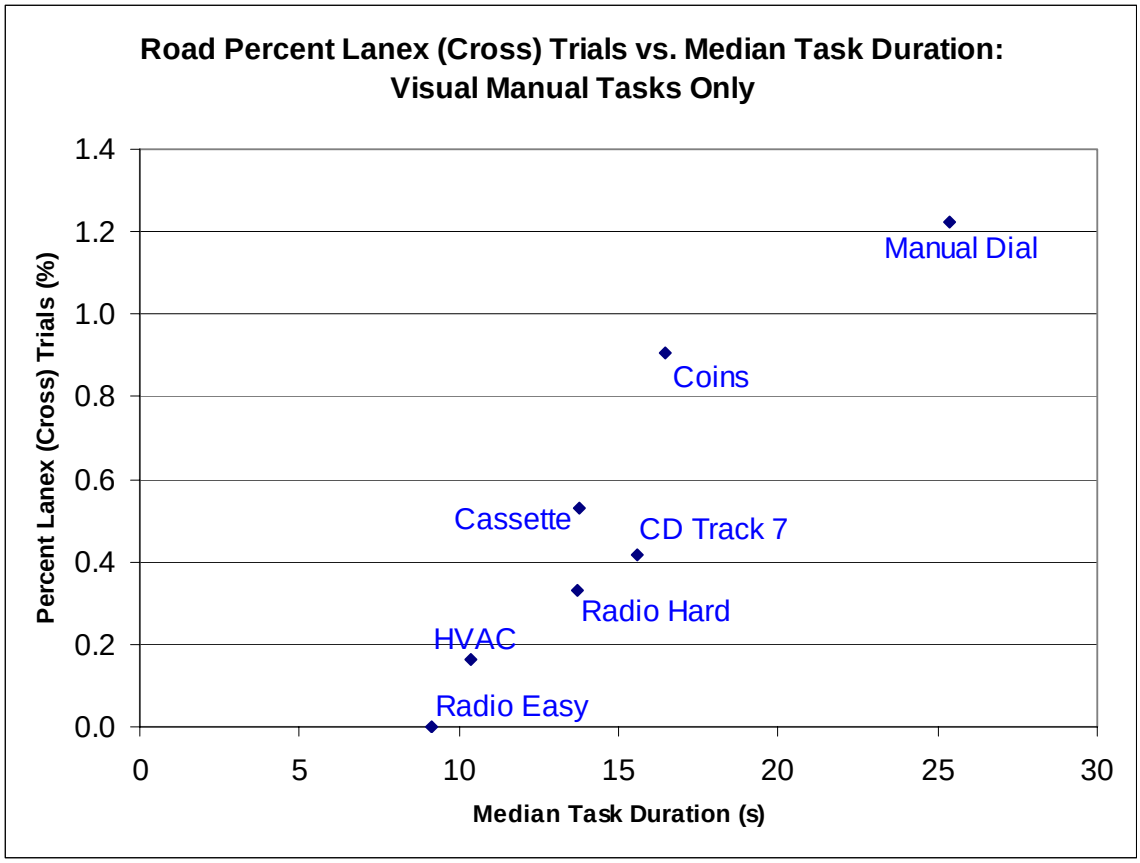


Figure 4-48. Road Percent Lanex (Cross) Trials by Task Duration for Visual-Manual Tasks Only

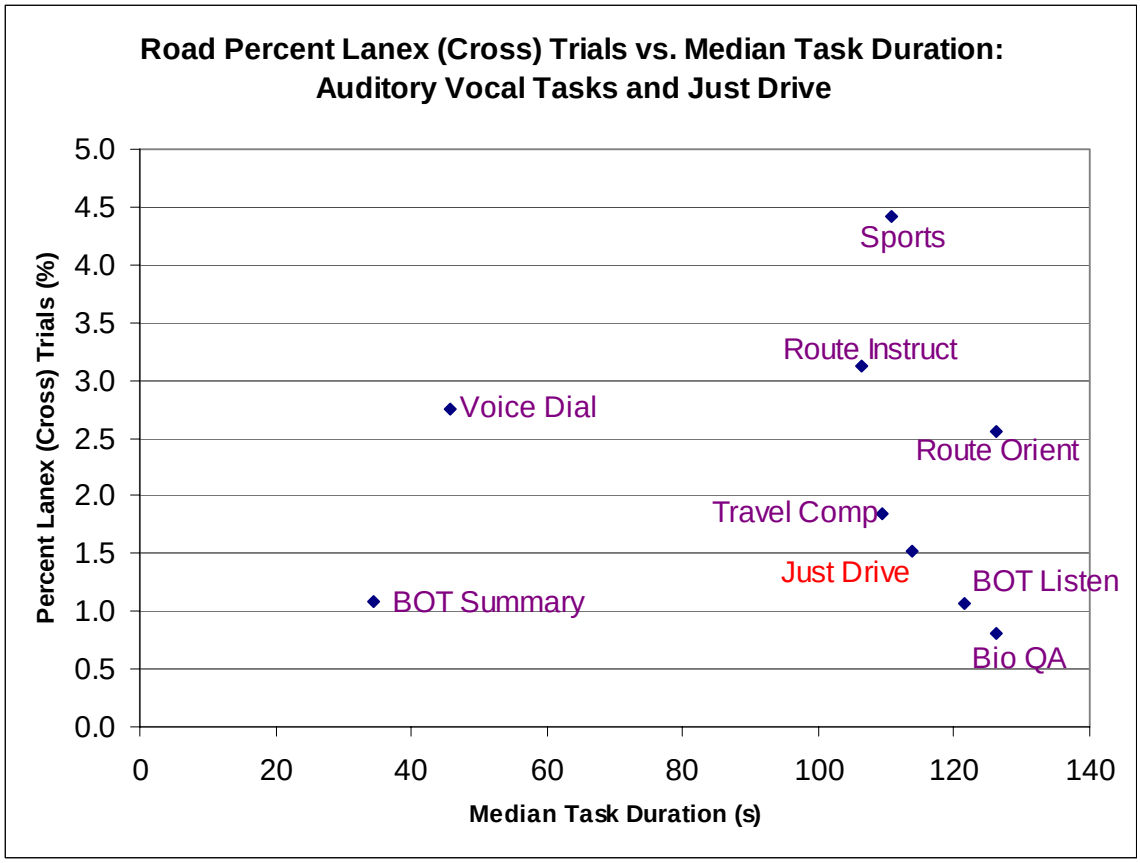


Figure 4-49. Road Percent Lanex (Cross) Trials Versus Task Duration for Auditory-Vocal Tasks and Just Drive

The repeatability of the lane exceedance measures (see Table 4-5) were assessed using the split group method previously described. The correlation between the Percent Lanex (Cross) values across tasks for the two groups on the road was approximately 0.76. Figure 4-50 is a scatter plot of the Group 0 and Group 1 data. Group 0 had more lane Exceeds than Group 1. This accounts for the considerable scatter between the two groups even though there is a positive relationship between the two groups. In either case, the percentages Lanex (Cross) trials was low and this might have prevented a better fit.

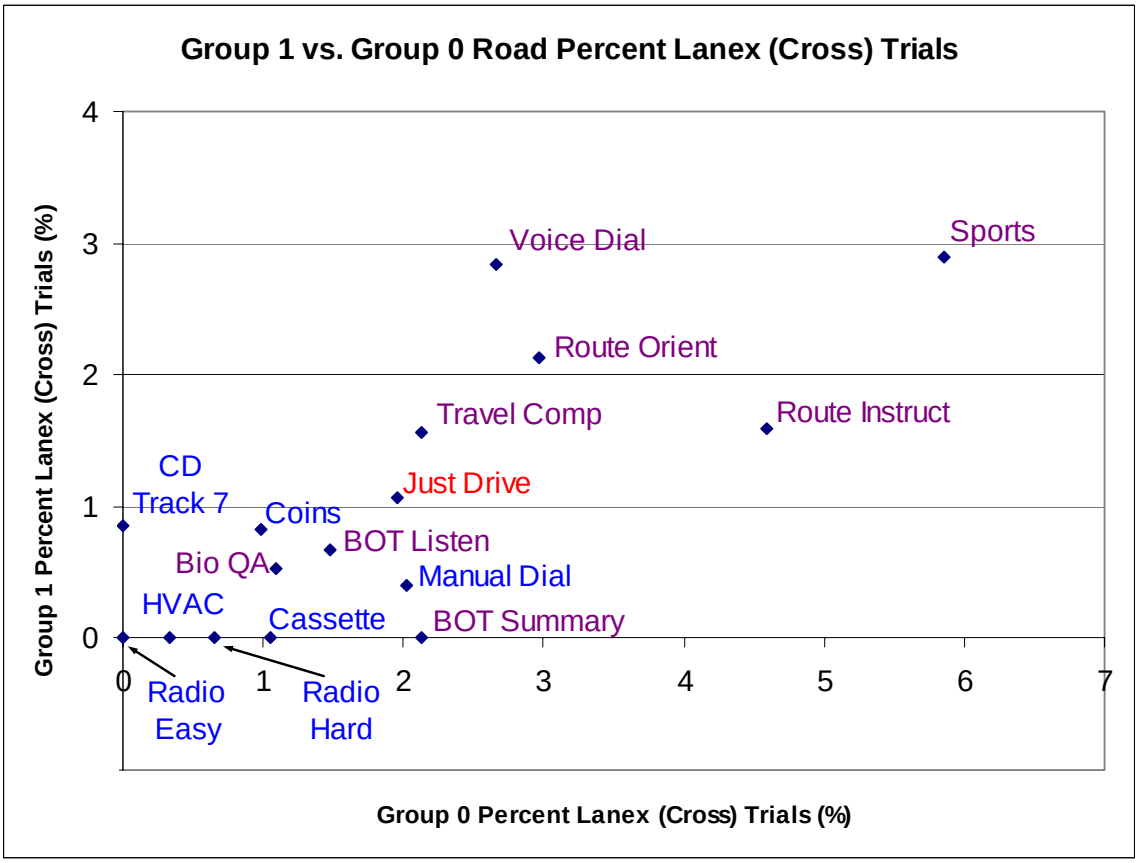


Figure 4-50. Group 0 Versus Group 1 Road Percent Lanex Cross Trials

A discriminability analysis was conducted on the road Percent Lanex (Cross) data for the visual-manual tasks and the auditory-vocal and Just Drive tasks separately. Of the six paired comparisons between Manual Dial and the remaining six tasks, four were statistically significant. Five of the six visual-manual tasks tested on the road were directionally correct with respect to prior prediction. However, the CD/Track 7 task did not have more Lanex (Cross) trials than the Manual Dial task. The reader should consider that these statistical results are based on very low incidence of Lane Exceeds. Most of the time, there were none.

A discriminability analysis was also done for the auditory-vocal and Just Drive tasks. Of 15 paired comparisons, only three were statistically significant ($p \leq 0.05$). Route Instructions and Route Orientation tasks were associated with a higher percentage of Lanex (Cross) Trials than Biographical Q&A. Route Orientation also had a higher percentage than Book-on-Tape Summarize. Seven of the 12 non-significant comparisons were directionally correct with respect to prior prediction. Again, the reader must consider these results in light of the very small number of lane exceed cases available.

4.5.3 Summary of Findings from Road Lateral Control Measures

The SDLP measure appeared to be applicable to the visual-manual tasks and less applicable to the auditory-vocal and Just Drive tasks. Four out of 6 paired comparisons between Manual Dial (a higher-workload task) and the remaining 6 lower-workload tasks were statistically significant all were directionally correct with respect to prior prediction. Four out of 6 paired comparisons with Percent Lanex (Cross) trials were statistically significant and 5 of the 6 comparisons were directionally correct. The latter results must be tempered by the very small numbers of Lanex events obtained.

The SDLP measure was not as applicable to the auditory-vocal and Just Drive tasks. Only 3 out of 15 paired comparisons were statistically significant. Nine of the 12 remaining comparisons were not directionally correct given prior prediction. Only 3 out of 15 paired comparisons of Percent Lanex (Cross) trials were statistically significant. Seven out of the remaining 12 non-significant difference were directionally correct with respect to prior prediction.

The SDLP and Lanex measures obtained in this study suggest that they were more applicable to the visual-manual tasks than to the auditory-vocal and Just Drive tasks. SDLP required a lane tracker to measure, with accompanying cost and complexity. Lane Exceedances were simpler to measure by direct observation of video recordings but were less frequent. Ideally, both types of data would be collected to assess driving performance.

4.6 Road Task Effects on Longitudinal Control

Longitudinal control is critical to maintaining vehicle separation. With degradation in longitudinal control, such as decreased range and increasing range rates, the potential for rear-end collision increases. Rear-end collisions account for a large number of accidents and systems to prevent dangerous ranges or closing rates are studied extensively. Thus, forward range and range rate are important metrics to examine for potential effects of driver distraction.

Another longitudinal metric is vehicle speed. Accidents caused by large variances in speed occur both in low visibility and dense traffic situations. Speed is also often a factor in run-off-road accidents that occur on curved sections of roadway.

Numerous measures of longitudinal control were examined in the DWM study. These measures include forward range, range rate, speed, and time headway. Measures of variance and central tendency such as minimum, mean, median, maximum, and standard deviations can be calculated. For this study, measures of range, range rate, and speed were selected for in-depth analysis.

In the DWM study, the vehicles driven by test participants were equipped with Delphi ACC1 forward range sensors. This sensor was modified to output information on the range in feet, range rate in feet per second, and lateral location of a vehicle ahead of the subject vehicle. Quality data was required for at least 90 percent of each individual task performance in order for any range data to be included in analysis. Speed of the subject vehicle was calculated from the OEM transmission sensor and recorded in feet per second.

For analysis of longitudinal measures, task performances were averaged across all replications of a task that did not contain a lead vehicle deceleration event for each participant. All tasks of a particular type, visual-manual, auditory-vocal, Just Drive and mixed-mode were then averaged across tasks and participants. For this analysis, the mixed-mode tasks, containing both visual-manual and auditory-vocal components were grouped separately into the mixed-mode task type. All Just-Drive tasks were also grouped separately from the other task types. These data were used as the input to analysis of variance to examine potentially significant task effects on vehicle control. While the results of this analysis are mentioned where appropriate, all graphs in this section present data by individual task. The task data were averaged across all tasks that did not contain a lead vehicle deceleration event for each participant. Data were then averaged across participants to yield a mean performance metric inclusive of all participants for each task.

4.6.1 Minimum, Mean, and Maximum Measures

Figure 4-51 shows the means of minimum, mean, and maximum range for each task with tasks ordered according to mean task duration for all on-road tasks. Table 4-10 presents the list of numeric codes used in Figure 4-51 cross referenced to task names. This table applies to other graphs also presented in this section. As was seen with the test track results, conventional visual-manual tasks are tightly clustered and show less variability in range than the longer auditory-vocal tasks. In between the two clusters are unique tasks, the shortest auditory-vocal task, Book-on-Tape Summarize, and the mixed-mode task Voice Dial. This graph again clearly shows the importance of task duration in regards to longitudinal vehicle control, similar to test track results.

Analysis of Variance for these measures showed statistically significant differences between tasks based on task type—visual-manual versus auditory-vocal. There was also some differentiation between tasks within the visual-manual grouping, however, these differences were small in magnitude. Interestingly, Task 19, Book-On-Tape Summary, lies between auditory-vocal and visual-manual tasks, as would be expected, if these differences are due mainly to task time. The other middle ground task is Voice Dial. While Voice Dial has components of both task types, it has a task duration between those types as well. This task falls between the other groups and thus lends further evidence to the importance of task duration for longitudinal control measures.

Figure 4-52 presents the (mean) minimum, mean, and (mean) maximum range values by task. The longer auditory-vocal tasks tend to have higher maximum range values and lower minimum range values than the shorter duration visual-manual tasks. The ordering of tasks here based on variation in range is similar to the ordering based strictly on task duration as shown earlier in Figure 4-42. The analysis of variance showed significant task-type effects for each of these three measures. The shorter visual-manual tasks have less variation in range than the longer auditory-vocal tasks with the Just Drive and mixed-mode task types ranking between the other two for variation in range.

Figure 4-53 contains mean measures of range rate and again shows the pattern of increasing variability with task duration. (See Table 4-10 for task names assigned to numeric codes.) This graph has tasks ranked by amount of variation in range rate and the order matches very closely to that obtained in ranking by task duration. Analysis of variance shows significant task effects between the shorter visual-manual and longer auditory-vocal tasks, with the Just Drive and mixed-mode task types grouping between the other two.

Figure 4-54 shows the variation in mean values of speed by task, again with visual-manual tasks having the least variation with auditory-vocal tasks having the most variation. There are statistically significant differences between the task types, especially visual-manual and auditory-vocal tasks with minimum and mean speeds. This appears to be an effect of task duration as the longer Just Drive and auditory-vocal task types are not significantly different while the shorter duration visual-manual and mixed-mode task types are different from the other types.

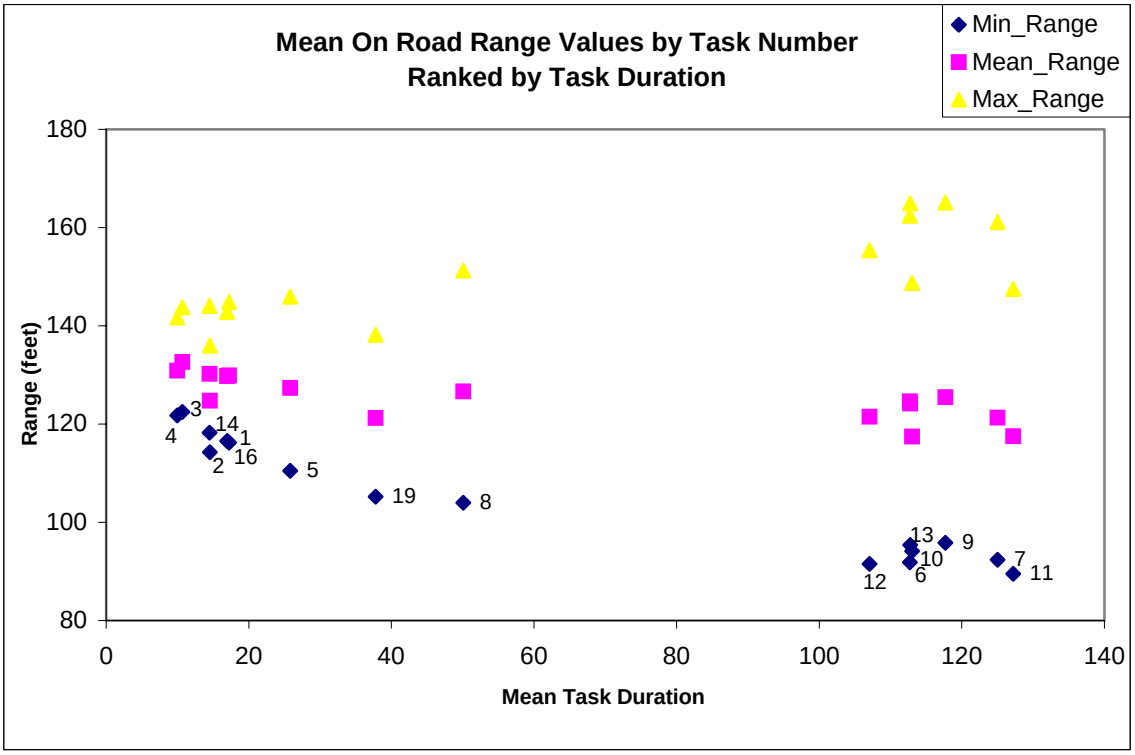


Figure 4-51. Mean On-Road Range Values by Task Number and Task Duration

Table 4-10. Numeric Codes Assigned to Tasks

Numeric Code	Task Name	Numeric Code	Task Name
1	Coins	13	Sports Broadcast
2	Cassette	14	Radio Tune Hard
3	HVAC	16	CD/Track 7
4	Radio Tune Easy	17	Route Tracing
5	Manual Dial	18	Delta Flightline
6	Travel Computations	19	Book-on-Tape Summary
7	Route Orientation	22	Destination Entry
8	Voice Dial	24	Read Text Easy
9	Book-on-Tape Listen	25	Read Text Hard
10	Just Drive	28	Read Map Easy
11	Biographical Q&A	29	Read Map Hard
12	Route Instructions		

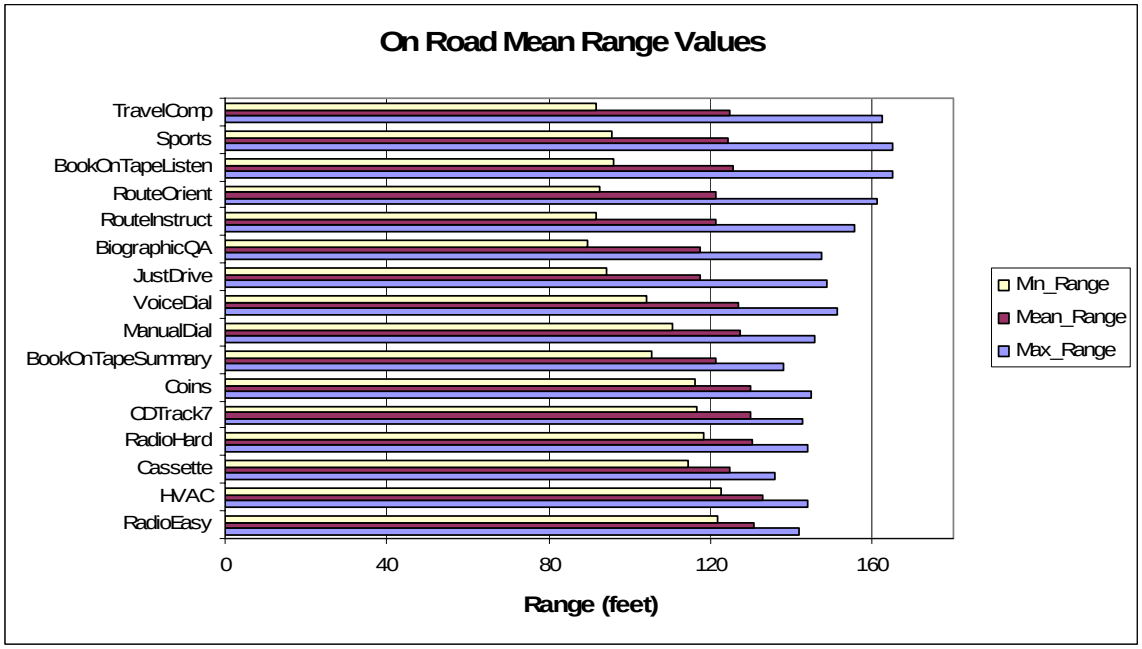


Figure 4-52. Mean On-Road Range Values by Task

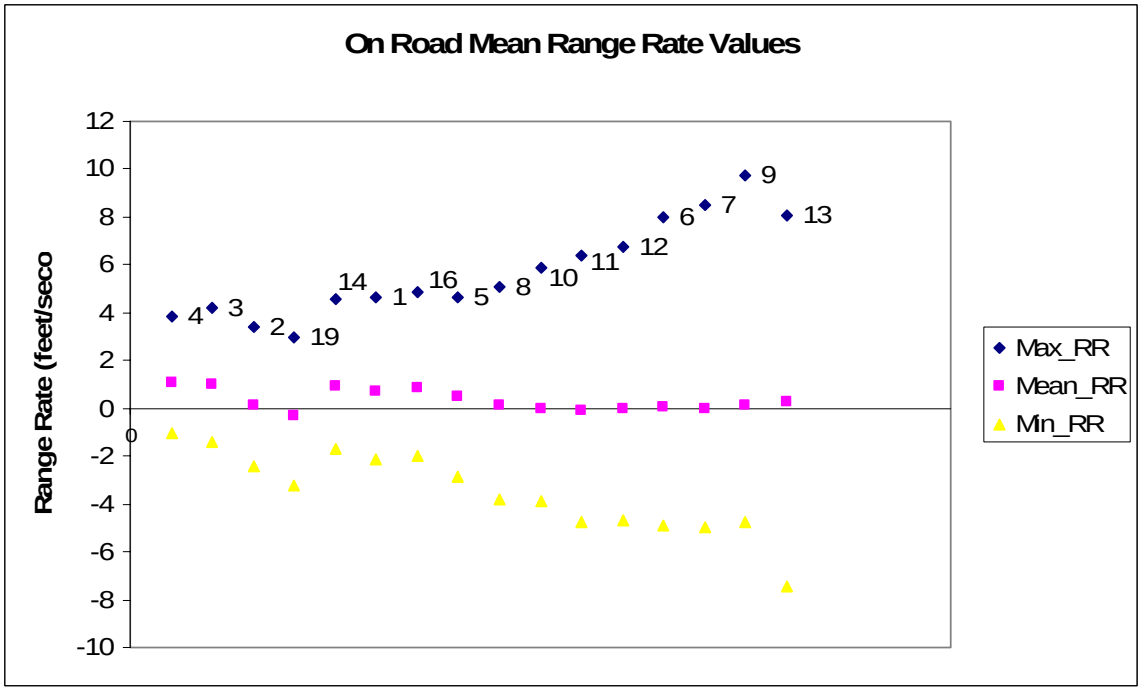


Figure 4-53. Mean On-Road Range Rate Values by Task

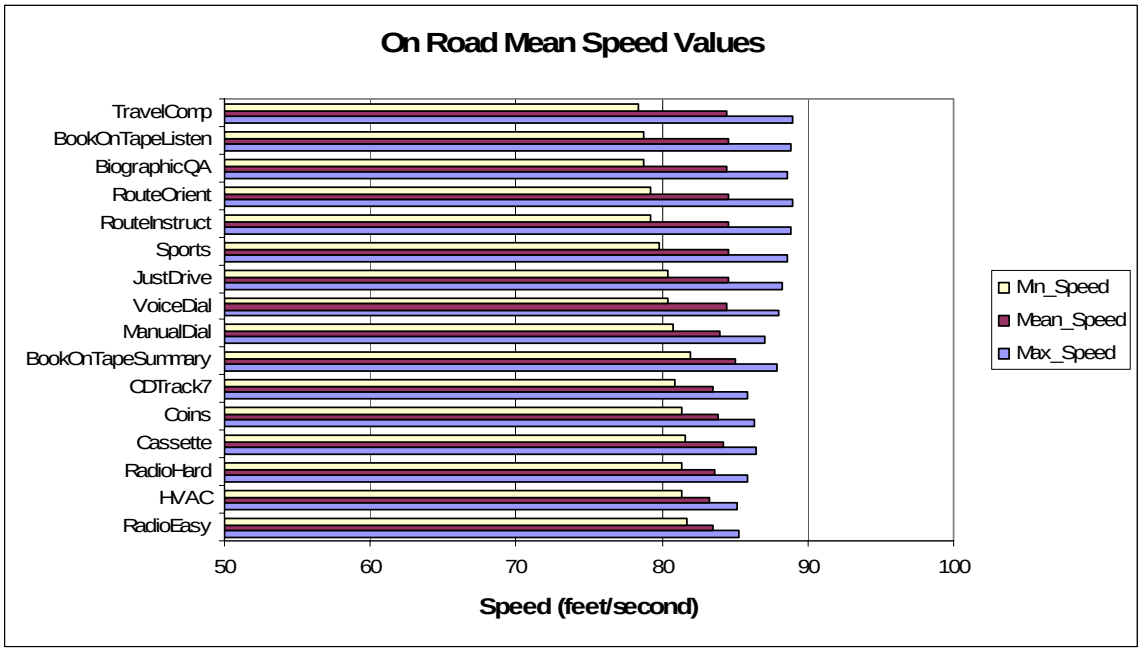


Figure 4-54. Mean On-Road Speed Values by Task

Figure 4-55 presents Speed Difference, computed as maximum minus minimum speeds, by task. The longer auditory-vocal tasks show the most variation in speed over the duration of the task. Auditory-vocal and visual-manual tasks are separated by Just Drive and the mixed-mode Voice Dial task. The short auditory-vocal task, Book-on-Tape-Summarize, has a speed difference very similar to the visual-manual tasks, which are closer in duration to this task than the other auditory-vocal tasks.

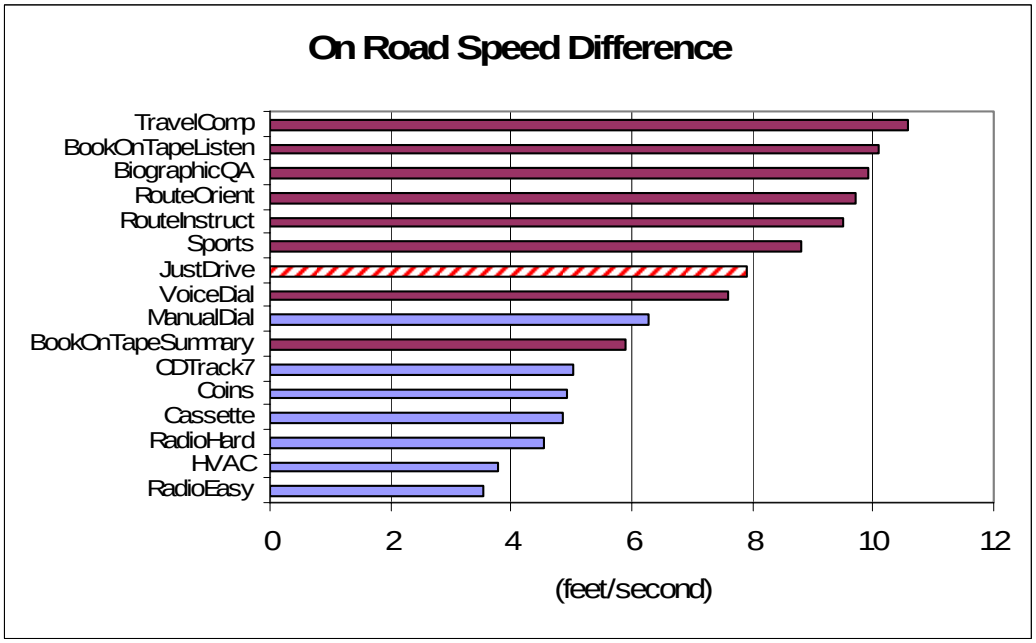


Figure 4-55. Mean On-Road Speed Difference (Max – Min) Values by Task

Figure 4-56 presents the relationship between Speed Difference and Task Duration by task. (See Table 4-10 for task names assigned to numeric codes.) The graph shows a very high correlation for the task set as a whole. This however, is due to the wide separation between conventional visual-manual tasks and the longer auditory-vocal tasks. Within the shorter visual-manual task grouping, correlation seems very high, with all tasks tightly clustered along a single regression line. The short auditory-vocal task, Book-on-Tape-Summarize, as well as the mixed-mode task, Voice Dial, lie between the two clusters of tasks. Both of these results indicate that task duration is important in longitudinal control. This however, cannot be said about the auditory-vocal tasks, which are much less tightly clustered. As with the test track results, this is an indication that task duration is not the sole influence on longitudinal control.

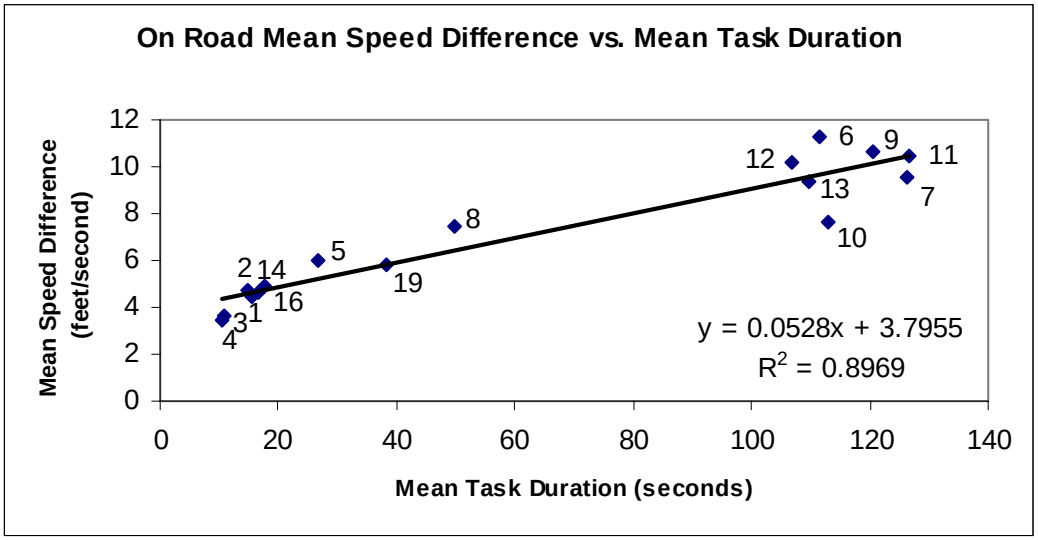


Figure 4-56. On Road Speed Difference Versus Task Duration

Figure 4-57 presents Speed Change computed as final minus initial speed of the test vehicle. While the differences in initial and final speeds are small, they display an interesting division of auditory-vocal and visual-manual task types, with Just Drive in the middle. When task types were examined with analysis of variance, both Speed Change and Speed Difference showed significant task effects. It is also interesting to note that Speed Change groups Book-on-Tape-Summarize with the auditory-vocal tasks instead of with the visual-manual tasks as Speed Difference does.

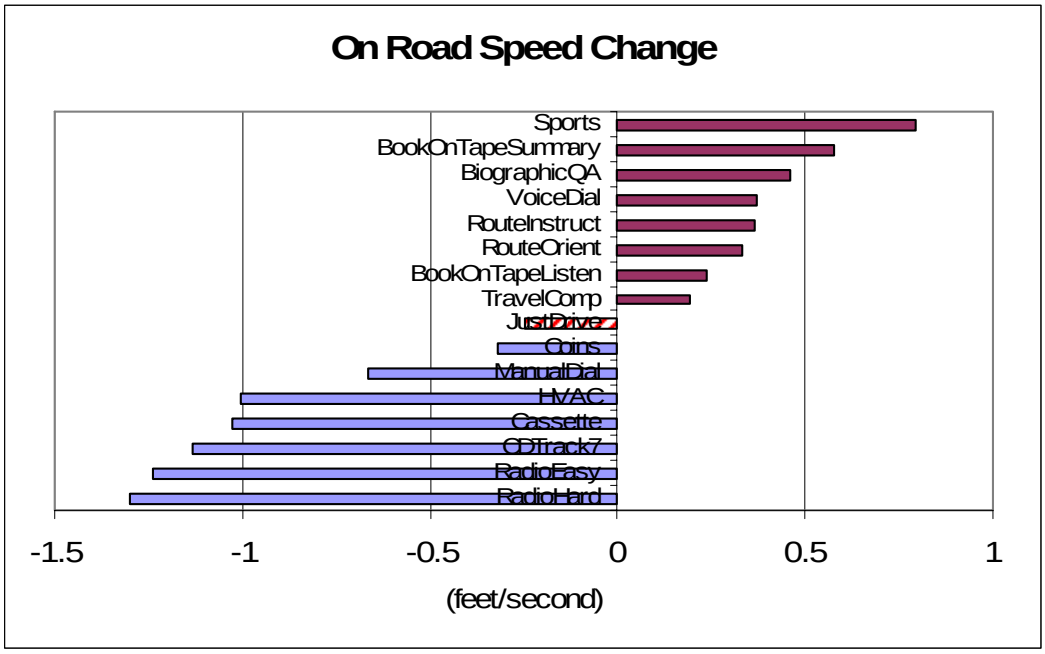


Figure 4-57. Mean On-Road Speed Change (Final – Initial) Values by Task

4.6.2 Split Group Reliability of Measures

To assess repeatability of longitudinal measures, a split-group repeatability analysis was performed. All on-road participants were divided into two groups of roughly equal distribution of age and gender. All measures were then summarized by task number and by task type within each group. High correlation between the same measures from the two groups was taken as evidence of repeatability of the metric.

Of all measures examined, metrics representing minimum or maximum values and standard deviations were the most repeatable, with all but one value (0.85) ranging from 0.95 to 0.99 for correlations between the two groups of participants. These are also the measures that are most correlated, with values typically ranging from 0.95 to 0.98, to mean task durations. Therefore, these were not considered further as possible measures of workload.

Repeatability of longitudinal measures was rather high, (see Table 4-11) with the most correlated measures typically being for extremes, minimums, and maximums. With correlations to mean task durations for all but mean range rate being ± 0.707 or better, yet lower than other correlations, these were the measures chosen for further analysis.

Table 4-11. On-Road Split-Group Reliability Correlations for Longitudinal Measures

On-Road Split Group Reliability		
Metric	**Task #**	**Task Type**
Min_Range	0.9808	0.9856
Mean_Range	0.8588	0.9732
Max_Range	0.9390	0.9821
Min_RR	0.7299	0.9509
Mean_RR	0.8451	0.9906
Max_RR	0.8921	0.7794
Min_Speed	0.9462	0.9549
Mean_Speed	0.9094	0.9080
Max_Speed	0.9860	0.9761
Speed_Difference	0.9853	0.9989
Speed_Change	0.9472	0.9640

Grouping tasks by type improves the repeatability of the measures of range and range rate, but lowers repeatability of speed measures slightly. This mix of improving or reducing correlation to task duration holds for all longitudinal measures

4.6.3 Summary of Findings from On-Road Longitudinal Metrics

In reviewing the on-road longitudinal data, the same trends appear as with test track data. When examining these measures by task or by task type multiple statistically significant effects can be found, and they tend to be the same effects that were significant for test track data. Rankings of tasks relative to one another tend to change somewhat between the two venues, however task groupings are very similar. In general, there is slightly more longitudinal variation with the on-road data, likely due to a less uniform roadway and heavier traffic in the on-road venue. These differences in variability combined with similar task durations may account for the changes of task ranks relative to one another.

As with the test track, due to the relative stability of automobiles longitudinally, changes in these measures are somewhat dependent on time. While mean values of range and range rate will be dependent on initial conditions and individual drivers' personal preferences, minimum and maximum values could again be a better indicator of the quality of longitudinal control. Similarly mean speed shows less variation between tasks and it is again more informative to examine Speed Difference and Speed Change. Range and range rate difference and change may also be informative, but were not examined in this study.

While longitudinal variation is somewhat dependent on time, due to vehicle dynamics, time does not explain all the task effects seen in these metrics. For instance, Speed Difference is correlated with task duration with an R^2 value of 0.90 for the entire task set in the on-road venue. This is again due mainly to two distinct groupings—short and long duration tasks. The correlations within each of these two groups indicate different relationships between time and speed between the task types. For visual-manual tasks, the correlation with task duration is very high and relatively similar to the test track results. The longer auditory-vocal tasks show much less correlation than the test track data. The increase in longitudinal variation for on-road data is at fault for this reduced correlation with the auditory-vocal tasks. The tasks are in the same relative

locations as they were for the test track data. These correlations, together with the results shown here, indicate that while important task duration is not the only influence on longitudinal control.

The summary data show that there are tasks with less variability in longitudinal position than Just Drive. These tend to be the short duration visual-manual tasks, which exhibit higher minimum and lower maximum measures of range, range rate, and speed. Longer auditory-vocal tasks tend to show more variability than Just Drive and in these comparisons, task durations are similar.

The first is that of a short-term “hold” occurring when a participant sheds the car following task to attend to the secondary in-vehicle task. Like steering, if a driver simply holds the accelerator pedal in the same position, for some short period of time the longitudinal position of the car relative to the lead vehicle will not change appreciably. After some time period however, numerous factors including friction, wind drag, and road surface will require an adjustment of the accelerator pedal to maintain longitudinal position. Thus, a short duration task with longitudinal variability significantly less than for just drive may be an indicator of a distracted state where a driver is not actively engaged in car following. A study of time series data and examination of accelerator pedal actuation and its relation to the longitudinal metrics may confirm this type of distraction.

The opposite condition, larger longitudinal variations, may be indicative of falling back. In this situation, which may start as a short-term hold, a driver is not actively attending to car following and begins to fall back from the lead vehicle. This is indicated by high positive range rates, increased range, and decreasing speed. At some point, the driver returns attention to the lead car and accelerates the vehicle to “catch up.” This is indicated by higher closing rates (negative range rate), decreased range, and increased speed. A time series study of driver behavior may confirm this condition is occurring by examining accelerator pedal position as well as the longitudinal metrics presented here. Major metrics of such an examination that may be useful would be frequency and amplitude of the variation in longitudinal position.

4.7 Comparisons between On-Road and Test Track Results

Track trials are, for some tasks, the only feasible venue for in-vehicle testing. It is important, because of this, to understand how comparable the results are between road and track. Comparability may differ measure by measure and be different on an absolute basis. Driver workload assessment, on the other hand, is not well defined. Given the current state of the art, some people consider it prudent to interpret workload measurements only in ordinal terms. This is why correlational analysis, across tasks, is used in this section.

Scatter plots of road versus track task-level selected summary statistics are provided in Figure 4-58 to Figure 4-68. It is apparent that only the tasks tested on the road could be compared with those also included in track testing. Potentially more demanding study tasks were only tested on the track and are excluded from this correlational analysis. The measures plotted cover selected measures of lateral control, longitudinal control, object-and-event detection performance, and selected measures of task-related driver eyeglance behavior. The description of each measure and observations on the correlation results are presented next.

4.7.1 Median Task Duration

Task duration is operationally defined as the elapsed time from task start, indicated by an experimenter button press after "Please begin now", to task end, indicated by an experimenter button press after the participant said "Done". The median of all participants is plotted with task duration measured in seconds. There is excellent agreement between road and track results as shown in Figure 4-58. This is to be expected for the auditory-vocal tasks since they were designed to last approximately two minutes (upper right-hand corner of the graph). However, median task durations are also highly similar for the visual-manual tasks whose duration were intrinsic to the tasks themselves. Those task durations were not fixed.

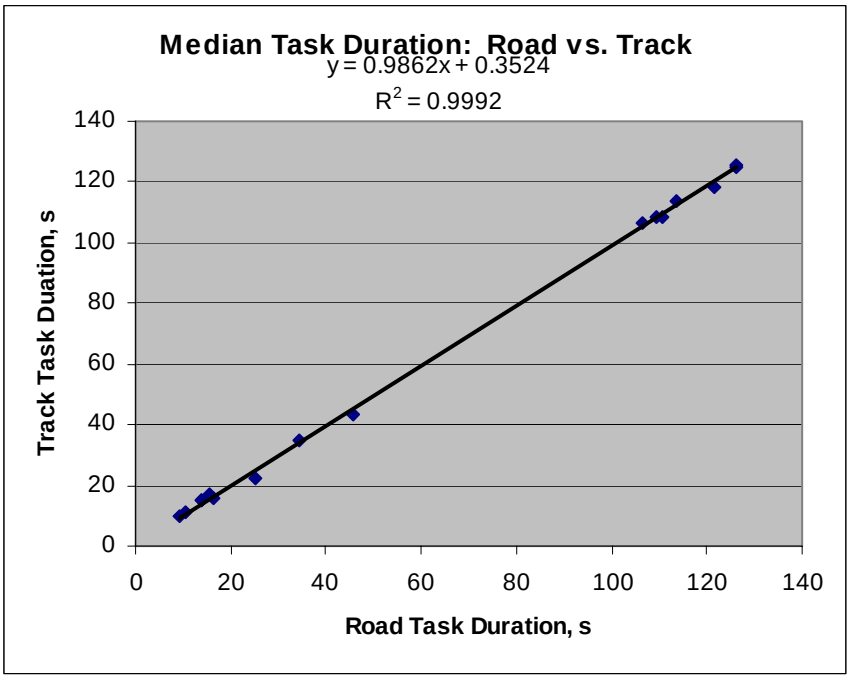


Figure 4-58. Correlation and Regression Between Road and Track, Median Task Duration

4.7.2 Median Standard Deviation of Lane Position

SDLP is the square root of the average square deviation in lane position about the mean lane position observed during the task duration. The median of all participants with valid lane position data is plotted. SDLP is measured in feet. SDLP correlation is high, though with some spread about the regression line. The range of SDLP for the road is slightly less than for the track as depicted in Figure 4-59. The assumption may be made that this range difference is due to the presence of other vehicles on the road. The clump in the upper right reflects the auditory-vocal tasks. These did not systematically impact SDLP results.

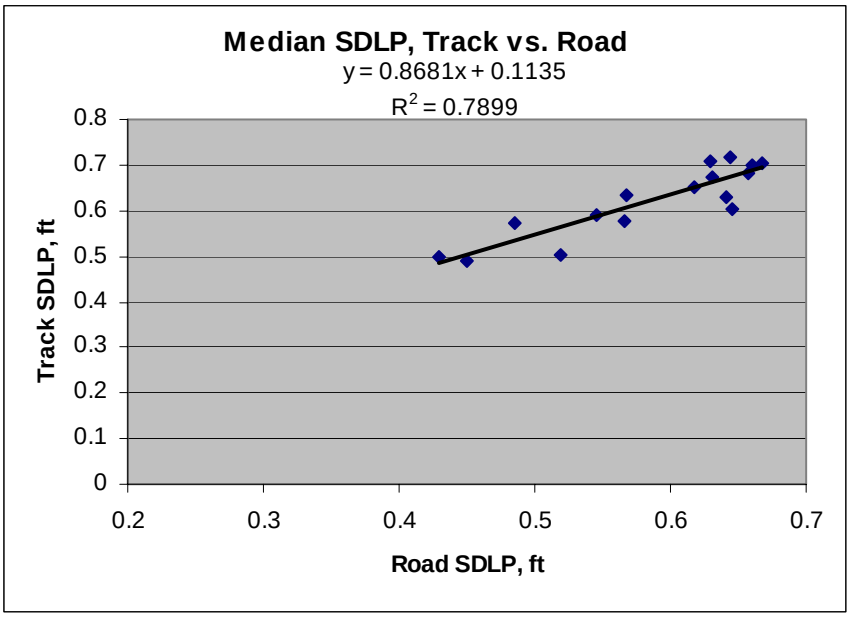


Figure 4-59. Correlation and Regression between Road and Track, Median SDLP

4.7.3 Percent Lane Exceedance (Cross) Trials

Percent Lane Exceedance (Cross) is operationally defined as the percentage of participants who had one or more lane exceedances during one or more task trials. A lane exceedance (cross) event was defined to have occurred if the leading edge of the participant vehicle crossed the adjacent lane line. This measure did not correlate between road and track (Figure 4-60). The test track was an oval and it was hypothesized that if these were removed, the correlation between road, which represented mostly straight road highway driving, and track would improve. It did not. The lack of a correlation might be due to the infrequent nature of lane exceedances, but this is hypothetical only and merits further research.

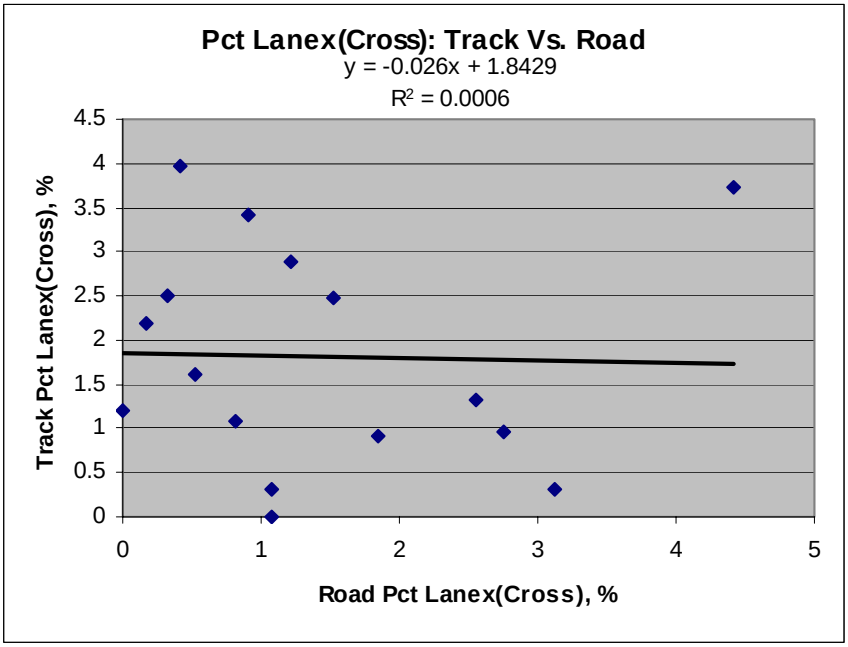


Figure 4-60. Correlation and Regression Between Road and Track, Percent Lanex (Cross) Trials (per participant, one or more events in a trial)

4.7.4 Median Speed Difference (SpeedDiff)

Median Speed Difference is defined in this study as the difference between the maximum speed and the minimum speed during the duration of a task. The median of all participants is plotted in ft/sec. The trend between road and track is clearly linear except for the curve at the upper right (see Figure 4-61). These points represent auditory-vocal tasks that did not vary systematically on this measure. For unknown reasons, there is a greater range of median values for tasks done on the road, compared to tasks done on the track (see Figure 3-51).

4.7.5 Percent LVD Miss Rate (LVDecel Miss Rate)

Percent LVD Miss Rate is the percentage of participants with one or more missed detections of a lead vehicle coast-down event during the task duration. It is measured in percentage points. The two points at the upper-right of the scatter plot in Figure 4-62 are for the two shortest tasks, HVAC and Radio (Easy). This OED stimulus was very difficult to stage with shorter tasks because of the inherent sluggishness of a vehicle coast-down. Furthermore, analysis revealed that the looming cue was often below published thresholds for detection (Mortimer, 1990), again due to the limited duration of shorter tasks. These results suggest that the LVD event must be considered carefully if evaluations of short tasks are contemplated.

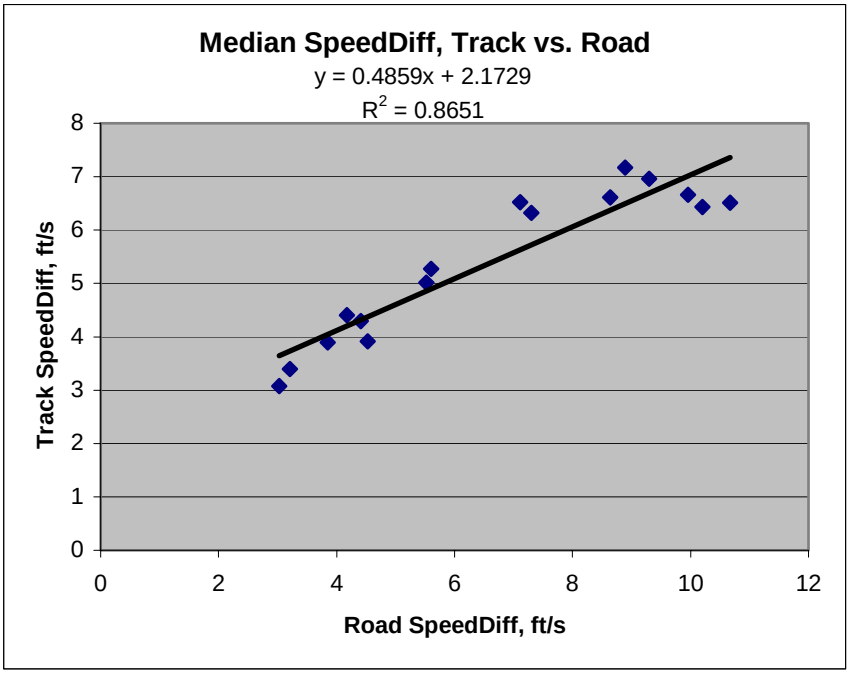


Figure 4-61. Correlation and Regression Between Road and Track, Median Speed Difference (Maximum – Minimum Speed during Task)

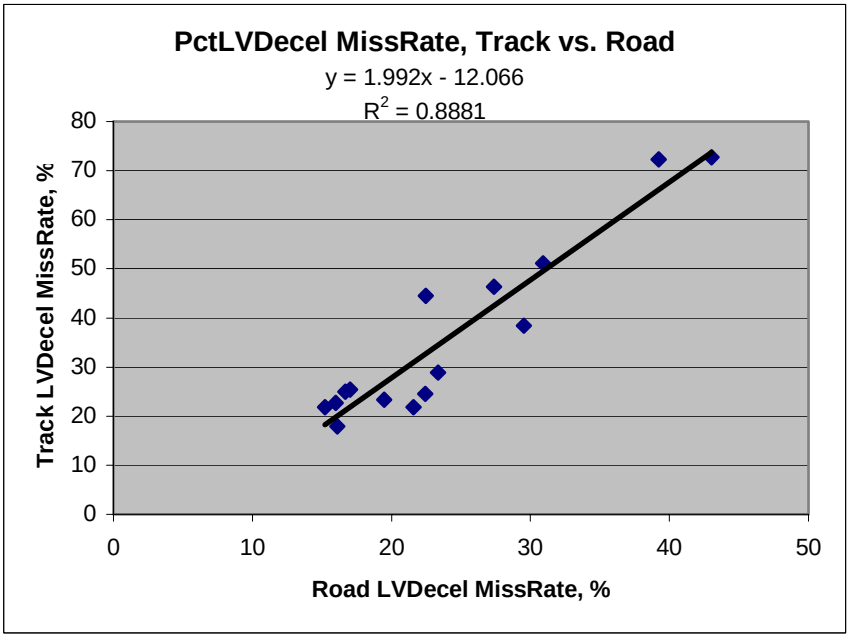


Figure 4-62. Correlation and Regression Between Road and Track, Percent LVD (LVDecel) Miss Rate

4.7.6 Percent CHMSL Miss Rate (CHMSL Miss Rate)

Percent CHMSL Miss Rate is the percentage of participants with one or more missed detections of a lead vehicle CHMSL whose duration was set to the instantaneous time headway at the time of CHMSL onset. It is measured in percentage points. The agreement between road and track results can be seen in Figure 4-63. The range of values is comparable between the two venues. A linear trend is apparent but there is a considerable amount of scatter about the regression line.

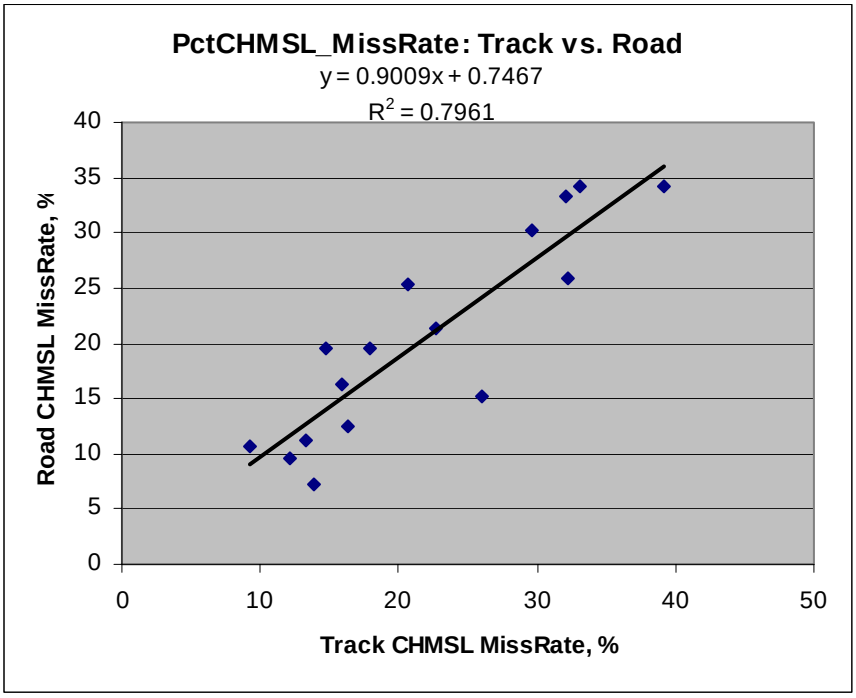


Figure 4-63. Correlation and Regression Between Road and Track, Percent CHMSL Miss Rate

4.7.7 Percent FVTS Miss Rate (FVTS Miss Rate)

Percent FVTS Miss Rate is the percentage of participants with one or more missed detections of a follow-vehicle turn-signal onset during the task duration. See Figure 4-64. There was, again, a positive correlation between road and track FVTS results.

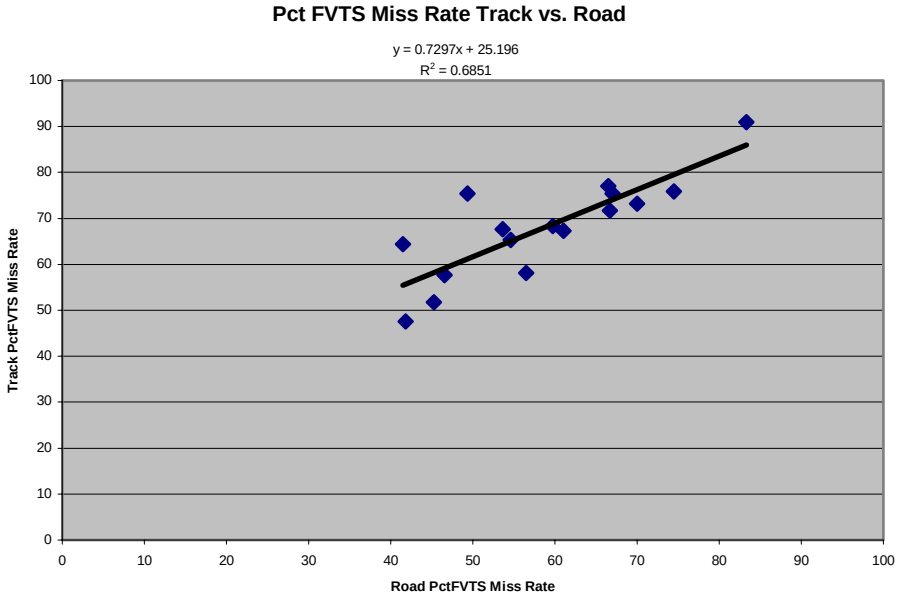


Figure 4-64. Correlation and Regression between Road and Track, Percent FVTS Miss Rate

4.7.8 Selected Eyeglance Behavior Measures

A number of Task-Related eyeglance measures were selected for comparison across the two venues. These measures are traditionally a part of driver eyeglance behavior assessment. (See other sections of this document for more comprehensive and in-depth analyses of eyeglance behavior on the road). These measures indicate excellent agreement between road and track trials for the DWM tasks assessed in both venues. Additional observations are provided below:

- **Average Task Related Total Eyes-off-Road Time** Total eyes-off-road time is operationally defined as the sum of durations for all task-related glances away from the road ahead, measured in seconds. Figure 4-65 indicates very good agreement between road and track results.
- **Task Related Percent Glance Time Away From Road** This is defined as the percentage of task duration spent on task-related glances away from the road ahead. The mean of available participant percentages obtained from road and from track trials are shown in Figure 4-66. There is good agreement between the two venues on this measure.
- **Average Number of Glances Away From Road** This measure is defined as the arithmetic average of the number of task-related glances away from the road ahead during a task trial. Results across road and track are comparable, as indicated in Figure 4-67. The correlation is quite high and the variability about the regression line is relatively small.
- **Average Task-Related Mean Single-Glance Time Away From Road** This measure is operationally defined as the arithmetic average over test participants of the mean duration of individual glances during a task trial, measured in seconds. Results across road and track are comparable as seen in Figure 4-68.

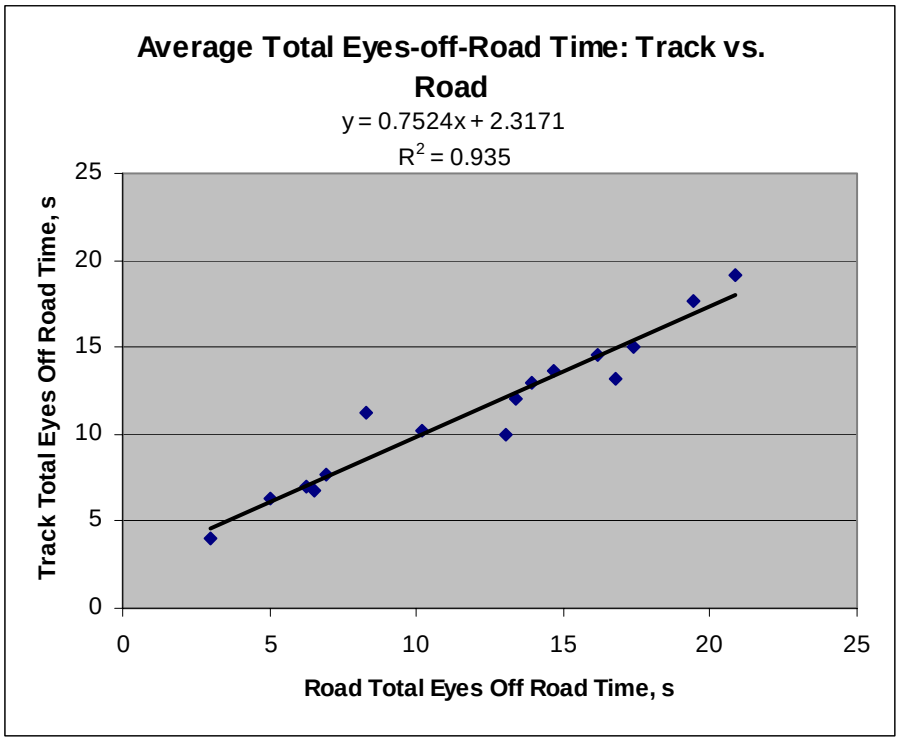


Figure 4-65. Correlation and Regression Between Road and Track, Mean Total Eyes-Off-Road Time

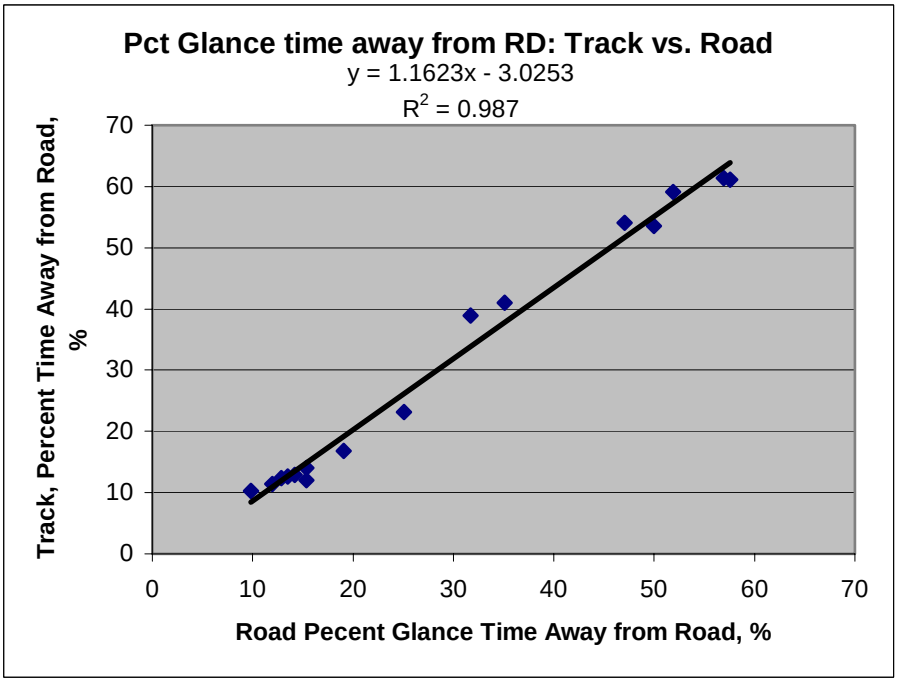


Figure 4-66. Correlation and Regression Between Road and Track, Percent Time Away from Road

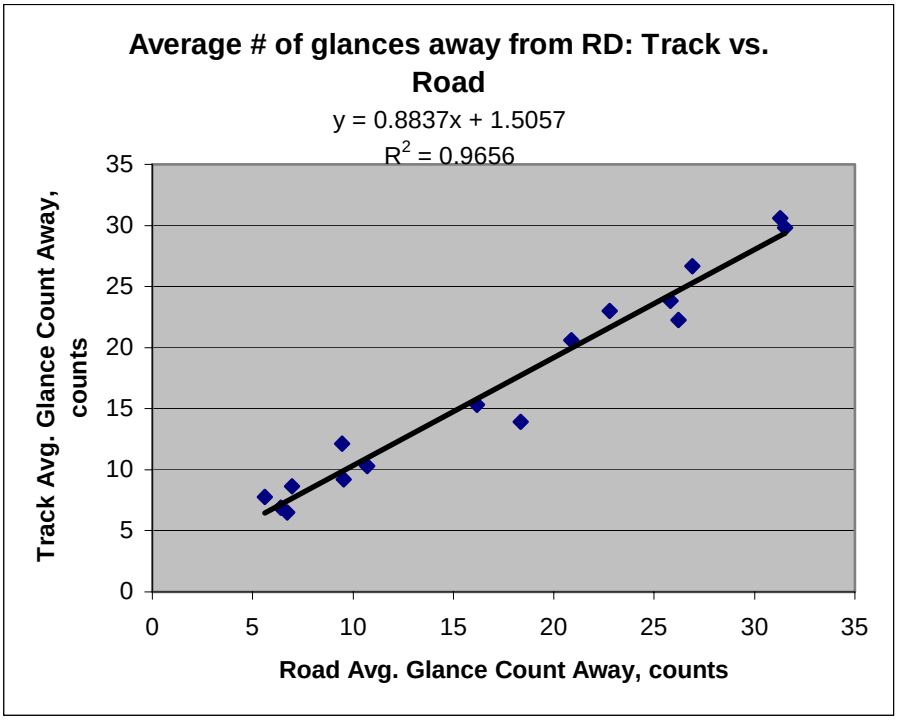


Figure 4-67. Correlation and Regression Between Road and Track, Mean Number of Glances Away from the Road

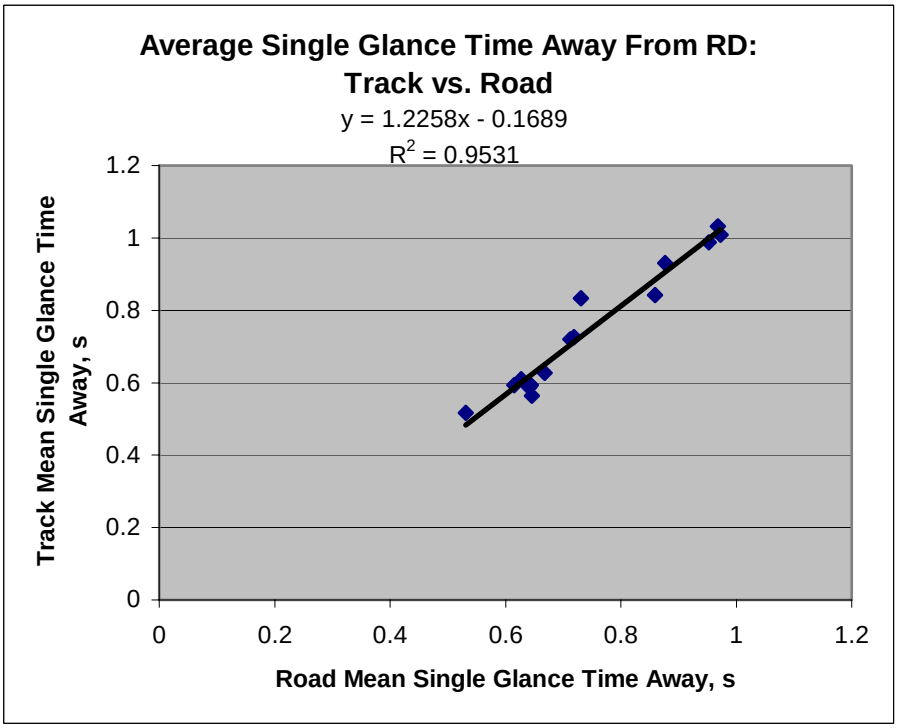


Figure 4-68. Correlation and Regression Between Road and Track, Average Single-Glance Time away from the Road

4.7.9 Summary of Comparisons between Road and Track Results

The road versus track correlation and regression plots largely speak for themselves. Similar trends across tasks are found in both venues with the exception of lane exceedances. The most consistent metrics were Task Duration and a few selected eyeglance measures. SDLP and SpeedDiff were subject to more variability about the regression line than the previous measures. The OED miss rates also showed more variability. Tests were conducted to determine whether this was due to the presence of curved sections on the oval test track. Removal of the lane exceedances that occurred on track curves, and subsequent comparison of data from the two venues, did not improve the correlation between road and track data. However, lane exceedances still are important indicators of degraded lane keeping. The lack of correlation between road and track for this variable may simply reflect the reduced range of task demands for the study tasks that were evaluated on the road.

4.8 Chapter References

Brown, I. D. (1994). Driver fatigue. *Human Factors*, *36*(2), 298-314.

Mortimer, R. (1990). Perceptual factors in rear-end crashes. *Proceedings of the Human Factors Society 34th Annual Meeting*, 591-594. Santa Monica, CA: Human Factors and Ergonomics Society.

Wierwille, W. W. (1993). Visual and manual demands of in-car controls and displays. In B. Peacock and W. Karwowski, W. (Eds.) (1993). *Automotive Ergonomics*. (pp. 299-320). London: Taylor and Francis.

5 Laboratory Results

Table 3-1 presents the distribution of laboratory test participants who contributed to this analysis. Six human performance methods were investigated in the laboratory. Three different subjective assessments of tasks also were investigated. Descriptions of these surrogates were provided in Chapter 2, *Study Design Overview*. Details of each test method are provided in appendices to this report. Table 5-2 summarizes the performance-based methods and describes the selected metrics analyzed from each. Table 5-3 summarizes the subjective assessment methods used and the metrics provided by them.

Laboratory results are presented in the following order:

- Surrogate measure repeatability
- Task effects of repeatable measures
- Discrimination of repeatable measures to distinguish higher-workload from lower-workload tasks
- Prediction by repeatable surrogate measures of selected driving performance and eyeglance measures

In general, repeatability and prediction were assessed through linear correlation and regression techniques. Discrimination of higher-workload from lower-workload tasks was conducted using the non-parametric sign test (Conover, 1999). An example of a more in-depth analysis, applied to Sternberg data, is provided in Appendix R. It illustrates how additional questions beyond task effects might be addressed with more sophisticated techniques.

Table 5-1. CAMP Driver Workload Metrics Laboratory Test Participants by Gender and Age Category

	Age Category						
	20's	30's	40's	50's	60's	70's	All
Male	4	5	5	3	5	4	26
Female	3	5	4	5	4	3	24
All	7	10	9	8	9	7	50

Table 5-2. Laboratory Tests Used in DWM Project

(Note: All measures collected during the duration of a task)

Method	Description	Selected Surrogate Measures	Applicability
Static Test Procedure	Person performed the requested task without any concurrent task or other interference. Completion time is collected.	Static Time, sec; Also known as Total Task Time-Static (TTTstatic), sec	Any task whose duration is "task-intrinsic," i.e., task ends when task goal is reached, not arbitrarily fixed
Occlusion Procedure	LCD goggles that open for 1.5 seconds and close for 2.0 seconds in a fixed cycle were worn as person attempted to complete a visual-manual task.	Total Shutter Open Time (TSOT), sec; the sum of goggle shutter-open times to complete a visual-manual task	Visual-manual tasks with task-intrinsic durations

Method	Description	Selected Surrogate Measures	Applicability
The R-Metric	At least two trials per person per task were needed, one Static trial and one Occlusion trial, to calculate the ratio R = TSOT/TTTstatic.	R-Metric, dimensionless ratio of TSOT/StaticTime	Visual-manual tasks with intrinsic task durations
Peripheral Detection Task –Alone (PDTA)	Single red laser light intermittently presented onto a blank screen in front of the participant while he or she was engaged in a task. Whenever the light stimulus was presented, the participant was to press a button.	Percent Missed Detections, %; Detect Response Time (RT)	In principle, all tasks
Peripheral Detection Task-in-STISIM (PDTS)	Same test procedure as with PDTA except stimuli were presented during an STISIM driving simulator (see next entry) run while the person was engaged in a task.	Percent Missed Detections, %; Detect RT	In principle, all tasks
STISIM	Part-task, fixed base driving simulator that required vehicle following at a nominal 55 mph on a highway with mostly straight road segments. A lead vehicle traveled at constant velocity. The participant completed tasks while trying to maintain a consistent, self-selected “comfortable” separation.	SpeedDifference (Max-Min Speed), ft/sec; Standard Deviation of Lane Position (SDLP), ft/sec; Lanex Percent Trials, %: percent of participants with one or more line crossings for one or more trials on a given task	All tasks
Sternberg-Spatial	Participant is given a set of three road signs (no wording, all geometric) to memorize. Then, during a task, probe road signs were presented on an LCD display. Task was to indicate whether the probe is part of the memorized set.	Percent Missed Detections, % Percent All Errors (Misses, Incorrect Responses), % Combined Decrement Score, dimensionless; the sum of Sternberg All Errors (as a proportion of all Sternberg trials), and an in-vehicle task performance score of 0.5 for a partially successful trial or 0.0 for a fully successful trial. Percent Error Given a Detection, % RT, Correct Responses RT, Incorrect Responses RT, All Responses	All tasks
Sternberg-Verbal	Same as Sternberg Spatial except that the road signs were route numbers rather than geometric road signs.	Same as above	All tasks

Many details on the surrogates are available in appendices to this report. These details include equipment specifications, stimulus materials, and test protocols.

Magnitude Estimation and the Sternberg Methodology deserve special mention by way of introduction. These are singled out because they may not be familiar to all readers. They were also exploratory in nature and used in a unique way in the DWM project.

Table 5-3. Subjective Workload Assessment Methods and Metrics Evaluated in the Laboratory

Method	Description	Metrics from Method	Applicability
Operator Workload (OWL)	A univariate subjective scale of task workload from 0 (low workload) to 100 (high workload). No anchor points (e.g., examples of low or high workload) are provided with this method. A task is rated against this scale.	OWL Rating: A dimensionless value on the OWL scale.	In principle, feasible for any type of task. May be applied to a description of tasks or after experience or trials with tasks. Just Drive not rated.
Multitasking Difficulty (MD) Magnitude Estimation	A rating scale of how hard it is to do a given task while driving and maintaining lane position, speed, headway, and detecting objects and events on or near the roadway). The test participant was presented a modulus or comparison task (Radio Tuning (Hard): "turn on the radio, switch to the FM band, and tune to a specific frequency."). This modulus was arbitrarily assigned a rating value of 100 by the experimenter. The test participant was asked to make his or her judgments about other stimuli (e.g., each other task) to reflect how many times greater or lesser a task's multitasking difficulty might be to the modulus task (i.e., estimate the ratio between the two stimuli, sensations, perceptions, or judgments). If a stimulus seemed twice as great as the standard, the test participant should say "200". If a stimulus appeared only half as great as the standard, the test participant should say "50" and so on.	Multitasking Magnitude Estimate: A dimensionless value calculated as the antilog of the mean log ratings for a given task.	In principle, feasible for all types of tasks. The modulus is given an arbitrary scale value of 100 and is not rated.
Situation Awareness Magnitude Estimation	A rating scale of how aware a test participant felt he or she was or would be to the roadway traffic and events while performing each task as compared to the standard task of "turning on the radio, switching to the FM band, and tuning to a specific frequency." All other details are similar to the magnitude estimation for multitasking difficulty.	Situation Awareness Magnitude Estimate: A dimensionless scale value calculated as the antilog of the mean log ratings per task	Feasible for all types of tasks. The modulus is given an arbitrary scale value of 100 and is not rated.

Magnitude Estimation is intended to produce a ratio scale (a scale of magnitudes, not just rank orders or equal intervals) of tasks placed along some psychological dimension such as Multitasking Difficulty or Situational Awareness. The procedure is based on the assumption that equal stimulus ratios produce equal subjective ratios. A test participant was presented with a standard for comparison. In the DWM project, the standard for comparison was a Radio (Hard) task. The test participant did not rate this task. It was instead arbitrarily assigned a scale value of 100 by the experimenter. A test participant was asked to make his or her judgment of the Multitasking Difficulty associated another task relative to the comparison task value of 100. Magnitude estimation is based on the premise that a test participant estimated the ratio on the Multitasking Difficulty dimension between a given task and the comparison task. If a task seemed twice as great as the standard on the psychological scale of Multitasking Difficulty, the test participant would say "200." If a task seemed only one-tenth as great as the standard, the test participant would say "10" and so on. The geometric mean of the numerical ratings provided by the sample of participants then provided the scale value on that psychological dimension. This method has been successfully and extensively applied to both physical stimuli (e.g., line length, brightness, loudness) and non-physical stimuli (social opinions, emotional stress) (Lodge, 1981; Gescheider, 1997).

Ratings, often used throughout the driver distraction and workload literature, can reflect task-set effects. That is, the ratings are likely to be influenced by the range of tasks that were part of the test. Ideally, an invariant scale (or set of scales) would allow researchers and evaluators at different times or in different locations to compare subjective assessments with confidence. One alternative to such an invariant scale would be to provide comparison tasks in each evaluation for interpretive guidance. In practice, this is difficult because OEM testing is highly constrained by time and other limited resources. Magnitude estimation might offer a solution to this problem, though it is not clear how it might differ from OWL or other subjective workload scales (e.g., NASA TLX, Hart and Wickens, 1990). The modulus or comparison task in magnitude estimation is a special case of behavioral anchors. Behavior anchors are descriptions assigned to different scale values to guide subjective assessments. In general, these are complex to develop with proper scaling properties (Fleishman and Quaintance, 1984). It is also necessary to have a good understanding of the anchors themselves.

The Sternberg Memory Search Method is well established in cognitive science as a means to study different cognitive processes (e.g., Sternberg, 1998). Sets of items (numbers, letters, pictures, etc.) are first memorized. Probes are presented and the participant responds "Yes" if the probe is a member of the memory set and "No" otherwise. Reaction times to different probes are used to infer the effects of various factors on performance. It has been used in a wide range of applied work as well. For example, Smith and Langolf (1981) used this method to assess the neuro-toxic effects of mercury exposure on industrial workers. By varying the set size, they were able to plot response time as a function of set size. The results showed an increased delay in responding. This effect was indicated by an increased slope of the line that related memory set size to response time, among the mercury-exposed workers relative to a control group. Wickens, Hyman, Dellinger, Taylor and Meador (1986) described various workload applications of the Sternberg memory search method. In particular, they described the use of Sternberg data to assess the workload of approach versus hold phases of flight. It has proven to be sensitive to a variety of workload effects.

The DWM project combined the Sternberg Method with both spatial and verbal road sign stimuli to examine dual task interference. However, the DWM project used only a single memory set size of three road signs per trial, all of the same kind (either spatial or alphanumeric). Traditional cognitive science research varies memory set size across trials. This was not done in the DWM project for practical reasons. Set sizes of four or larger appeared too difficult to apply to a wide

range of test participants. There was also not enough time to run additional trials with other set sizes. Wickens et al., (1986), aware of this type of constraint, also recommended a single memory set size of three or four items to provide stable workload measures sensitive to all stages of task demand. Analyses were conducted on visual-manual tasks separate from auditory-vocal and Just Drive tasks. When possible, analyses were conducted that made use of all tasks

The results will be presented within an analysis in sets. These sets include subjective assessments taken together; static time, TSOT, and the R-Metric taken together; STISIM driving surrogate measures taken together; and object-and-event detection (OED) measures taken together. The rationale for each grouping is provided next.

Subjective workload ratings of DWM task workload are presented together for the following reasons. They can be collected without additional apparatus or instrumentation. They can be applied to all subsidiary tasks. They do not involve task performance, though they may be collected during or after task performance. They reflect subjective impressions of workload that may differ from task performance effects (Hart and Wickens, 1990).

Static Time, TSOT, and the R-Metric are presented together for several reasons. They do not involve dual-task performance. Two of the three metrics, TSOT and the R-Metric, are appropriate only for visual-manual tasks. One of the three measures, the R-Metric, is defined in terms of the other two (R-Metric = TSOT/Static Time). And criteria have been proposed for all three (SAE, 2004; Alliance of Automotive Manufacturers, 2003; Asoh, Uno, Noguchi, and Kawasaki, 2002; ISO, 2004).

The Systems Technology, Inc., part-task simulator (STISIM Drive) produced several surrogate measures of driving performance. Standard Deviation of Lane Position (SDLP) and Percent Lanex Cross trials were chosen as lateral control measures. SDLP provided a continuous measure of lanekeeping while Percent Lanex Cross Trials provides discrete lane keeping violations. Speed Difference (SpeedDiff), i.e., the difference between the maximum and minimum speed during a task, was taken as a longitudinal control measure. The STISIM scenario involved car following at a self-selected “comfortable” following distance. The lead vehicle traveled at constant speed. The test participant was instructed that driving was the most important task and subsidiary tasks were to be performed only if and when the participant thought it appropriate to do so. The SpeedDiff was chosen as a robust measure that would reflect variations in range, range-rate, and speed change between the start and end of a task.

Several different OED methods produced missed detection data as well as detection latencies or response times. These measures and methods may tap into different aspects of distraction than subjective assessments, static time measures, or driving performance measures. Thus, the OED measures were grouped together for presentation purposes.

5.1 Repeatability Results

Surrogate measure repeatability was assessed using linear correlation and regression techniques. Two sub-samples of participants, of approximately the same distribution of age category and gender, were created through random assignment. The two subgroups' performance measures were summarized and compared statistically to assess the consistency of rankings across tasks at the level of task summary statistics (median Total Shutter Open Time; mean Operator Workload or OWL ratings, etc.).

A preliminary correlation and regression analysis of various surrogate measures was carried out, per task, to compare test-retest reliability. Test-retest reliability assessed the consistency of responses among the same participants who performed the same task twice. The correlation of the two repetition trials (reps) was generally very low. This occurred despite responses obtained from

the same person, who performed the same task twice, with the same equipment or materials, within a one-day or two-day period. Such results have been reported in the literature, albeit with more complex tasks (Lane, Kennedy, and Jones, 1986).

To improve the repeatability of measures, several steps were taken. The reps per test participant were averaged into a single number per person. Data from only one rep was used if that was the only data available. From these data, summary statistics were then calculated per task. Repeatability analysis then assessed consistency in task ranks or ordering across task summary statistics. For example, the mean of Operator Workload (OWL) scores per task was calculated using the data from N participant ratings. This was done separately for each of the two subgroups of task participants. Then the 22 tasks (Just Drive was not assessed) were correlated across Group 0 and Group 1. That is, each task had two mean OWL scores, one from Group 0 and one from Group 1. A correlation was then calculated across the 22 mean OWL scores.

The Split-Group repeatability results for selected surrogates are presented in Table 5-4. Repeatability correlation of $r = 0.70$ was used as an operational definition of a repeatable measure. This value was chosen to reflect correlations that account for approximately 50 percent of the variance in the data. This is reflected in the R-squared values included in the table. Ghosted entries are non-repeatable measures, as defined here.

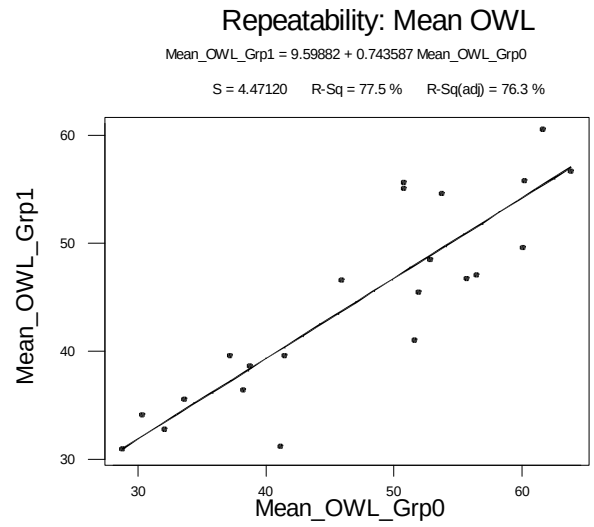
Table 5-4 indicates the selected surrogates generally had high split-group repeatability. The repeatability correlation coefficients were generally well above the $r = 0.7$ criterion. All those reported above the criterion were statistically significant at the 0.05 alpha level or beyond. This provided a variety of surrogate methods suitable for further analysis.

Group 0 versus Group 1 scatter plots were prepared for each repeatable surrogate measure to gain further insights into measure repeatability. These scatter plots differ in the measures presented. This makes it difficult to make direct comparisons across measures. However, variability about the regression line provides a visual indication of the correspondence between the two groups' results.

Figure 3-1 presents the repeatability scatter plots for OWL and Multitasking Difficulty measures. Linear relationships across the two groups are apparent. However, variability about the fitted line was moderately high for each. This variability reflects participant differences, measurement noise, or both inherent in the subjective assessments.

Table 5-4. Split-Group Repeatability for Selected DWM Laboratory Metrics

Lab Surrogate Measure	Split GroupTask Level Correlation, r	Split Group R^2 %
Mean OWL	0.880	77.5
Situational Assessment (SA) Scale Value	0.392	15.4
Multi-Tasking Difficulty (MD) Scale Value	0.882	77.9
Mean Static Time	0.996	99.3
Median Static Time	0.995	99.0
Mean Total Shutter Open Time (TSOT)	0.996	99.2
Median Total Shutter Open Time (TSOT)	0.997	99.5
Mean R_Metric	0.865	74.8
Median R_Metric	0.816	66.5
Mean STISIM Task Duration	0.998	99.6
Median STISIM Task Duration	0.998	99.5
Mean STISIM SDLP	0.951	90.4
Median STISIM SDLP	0.919	84.4
STISIM Percent Lanex Cross Trials	0.960	92.2
Mean STISIM Speed Diff	0.938	88.0
Median STISIM Speed Diff	0.911	83.0
PDT Alone Miss Rate	0.839	70.4
PDT Alone Mean of Mean RT	0.923	85.3
PDT Alone Median of Mean RT	0.881	77.7
PDT in STISIM Miss Rate	0.860	74.0
PDT-in-STISIM Mean of Mean RT	0.918	84.3
PDT-in-STISIM Median of Mean RT	0.773	59.7
Sternberg Percent Missed Detections	0.976	95.3
Sternberg Error Given Detection	0.240	5.8
Sternberg Percent All Errors	0.956	91.4
Sternberg Mean Correct RT	0.850	72.2
Sternberg Median Correct RT	0.871	75.8
Sternberg Mean Incorrect RT	0.204	4.2
Sternberg Median Incorrect RT	0.209	4.4
Sternberg Mean All RT	0.834	69.5
Sternberg Median All RT	0.855	73.1
Sternberg Combined Decrement Score	0.955	91.3



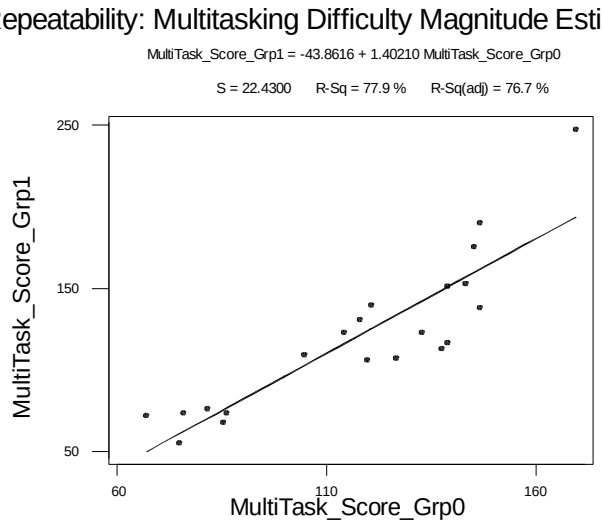


Figure 5-1. Repeatability: Group 0 Versus Group 1 Operator Workload and Multitasking Difficulty Ratings

Much tighter correspondence between Group 0 and Group 1 was found in the median Static Time and median Total Shutter Open Time (TSOT) measures presented in Figure 5-2. The difference between the mean and median values for these measures was quite small. On the other hand, the mean and median R-Metric plots showed important differences. The means exhibited extreme values on either side of an amorphous group of tasks in the middle range. These extreme values contributed to the relatively high correlation. The median R-Metric plot showed a better spread about the regression line but relatively high variability about the regression line remains.

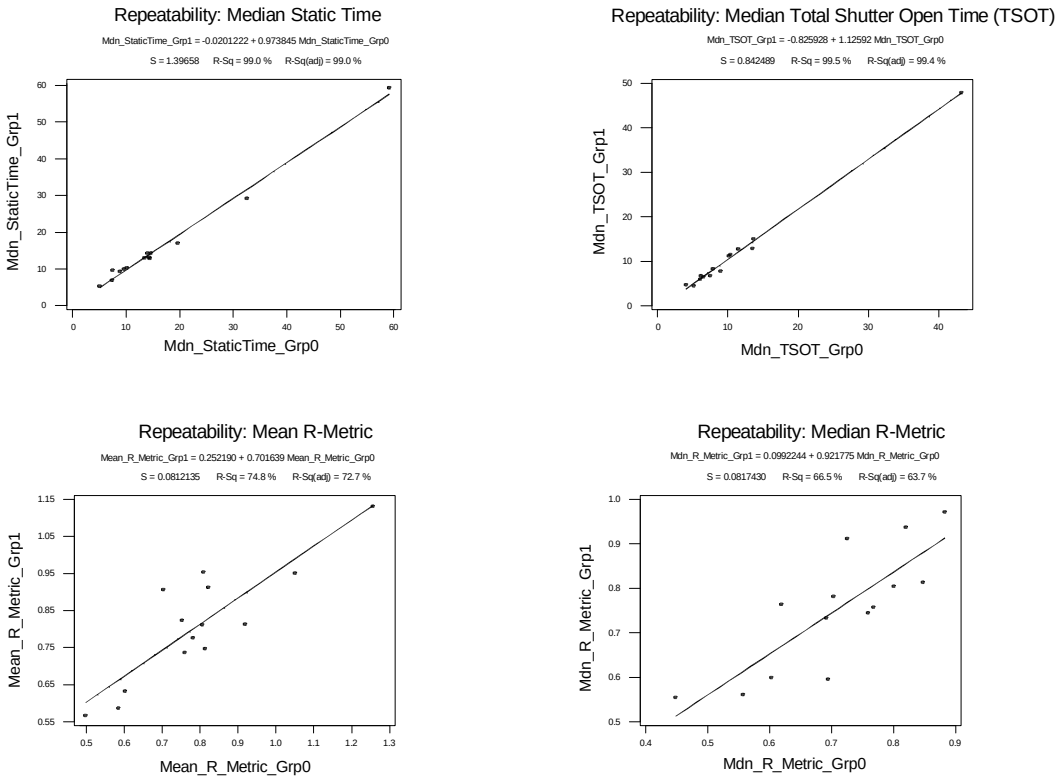


Figure 5-2. Repeatability: Group 0 Versus Group 1 Static Time, TSOT, and R-Metric Summary Statistics

A variety of STISIM measures were selected as driver workload surrogates. Figure 5-3 presents the repeatability scatter plots for mean and median STISIM Task Duration and Standard Deviation of Lane Position, and in the STISIM Percent Lanex (Cross) Trials. The mean SDLP plot was tighter about the regression line than the median SDLP plot. This pattern was opposite for that found with the R-Metric. The percent Lanex (Cross) trials showed relatively small variation about the regression line. The data also showed a curvilinear pattern whose origin remains unknown. No complex curve fitting was attempted for such data because of the relatively small number of data points and the interpretive difficulties that would ensue.

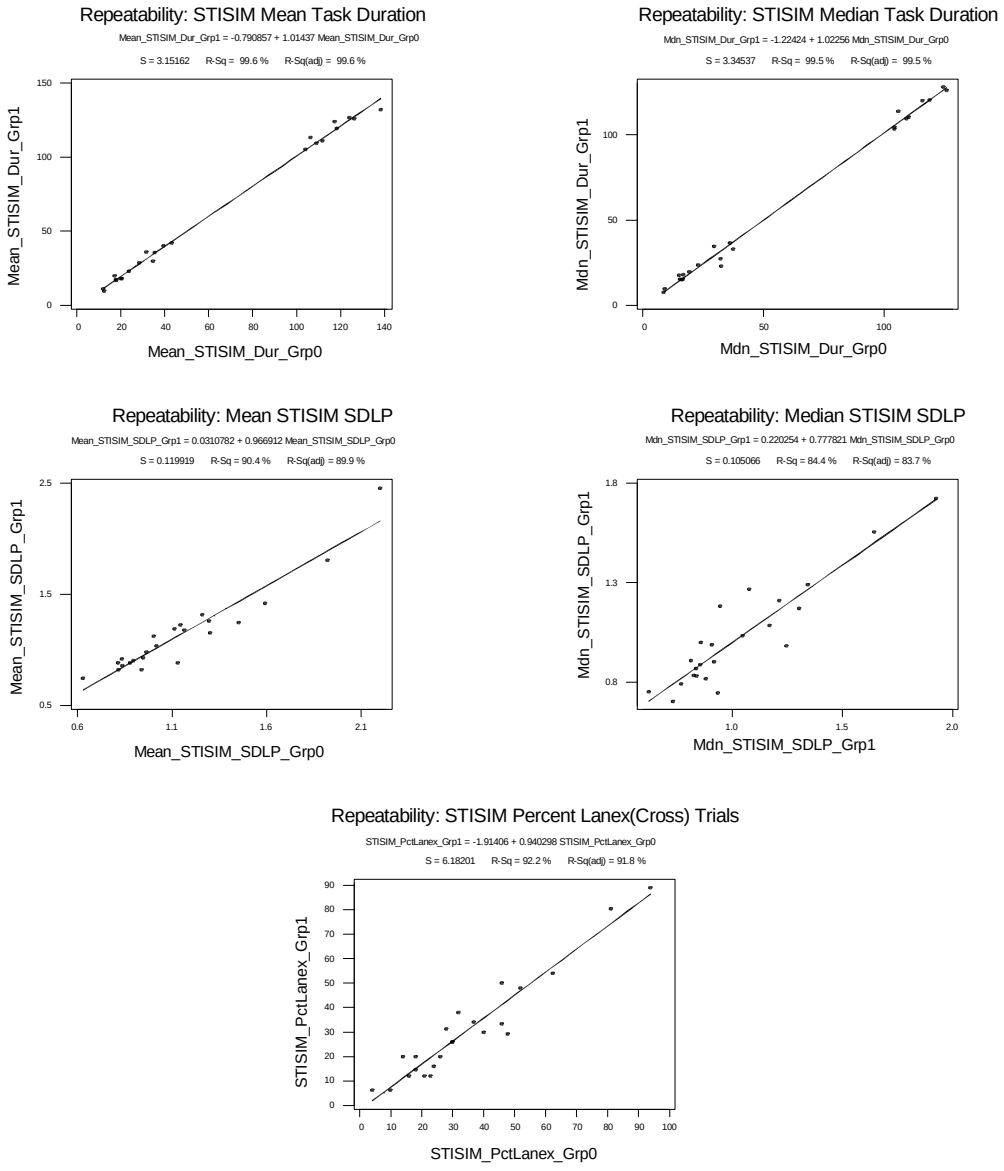


Figure 5-3. Repeatability: Group 0 Versus Group 1 Scatter Plots for Selected STISIM Measures

STISIM Speed Difference (SpeedDiff) repeatability plots are presented in Figure 5-4. Repeatability is higher for the mean values than for the median SDLP values. The standard deviation about the fitted line was approximately the same for either measure.

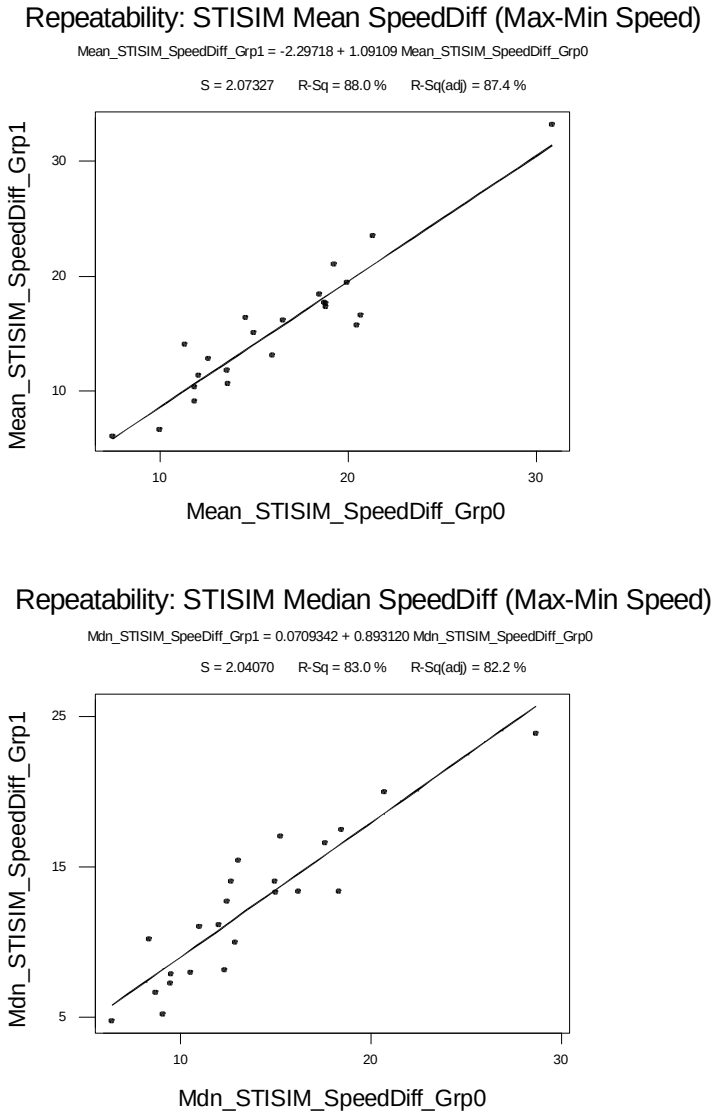


Figure 5-4. Repeatability: Group 0 Versus Group 1 STISIM Speed Difference Summary Statistics

A variety of OED miss rates are presented in Figure 5-5. An interesting pattern emerges as one moves from the PDTA to PDTS to Sternberg measures. The correspondence or repeatability improves as one moves from a simple detection and response task (PDTA) to that same task performed while driving (PDTS), to a more complex detection task that required memory scanning and a choice response (Sternberg). This suggests that increased load imposed while concurrently performing tasks reduced variability across groups of performers. Increased consistency appears to come with more demands placed on the performer that constrained performance

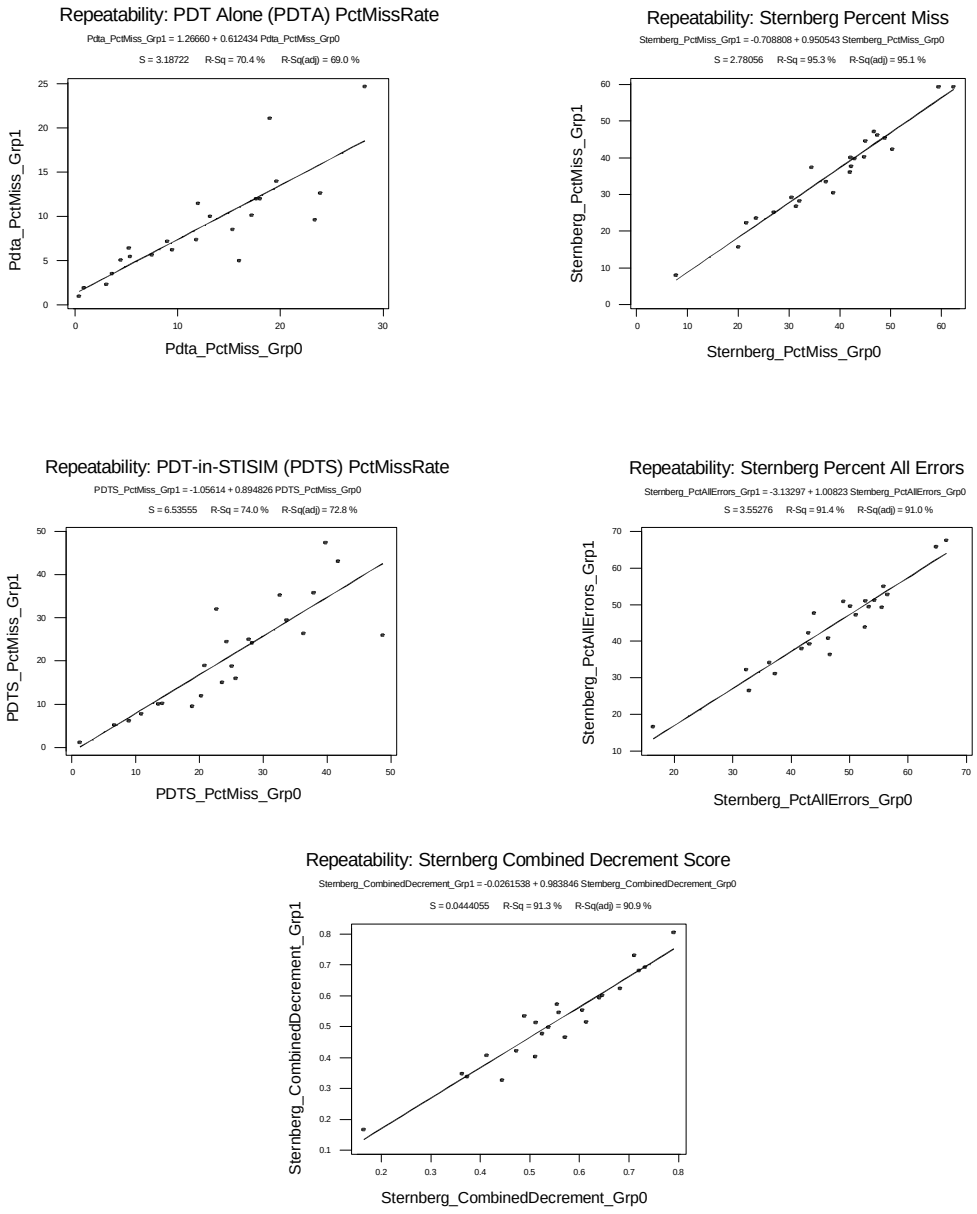


Figure 5-5. OED Repeatability: Group 0 Versus Group 1 PDTA, PDTS, Sternberg Percent Miss, Sternberg Percent All Errors, and Sternberg Combined Decrement Scores

The final set of repeatability plots presented in Figure 5-6 are for various OED Response Time (RT) measures. Higher miss rates were generally correlated with longer response times. Mean RT values for PDTA and PDTS had smaller variation about the fitted line than with Median values. On the other hand, there were virtually no differences between mean and median Sternberg RT values.

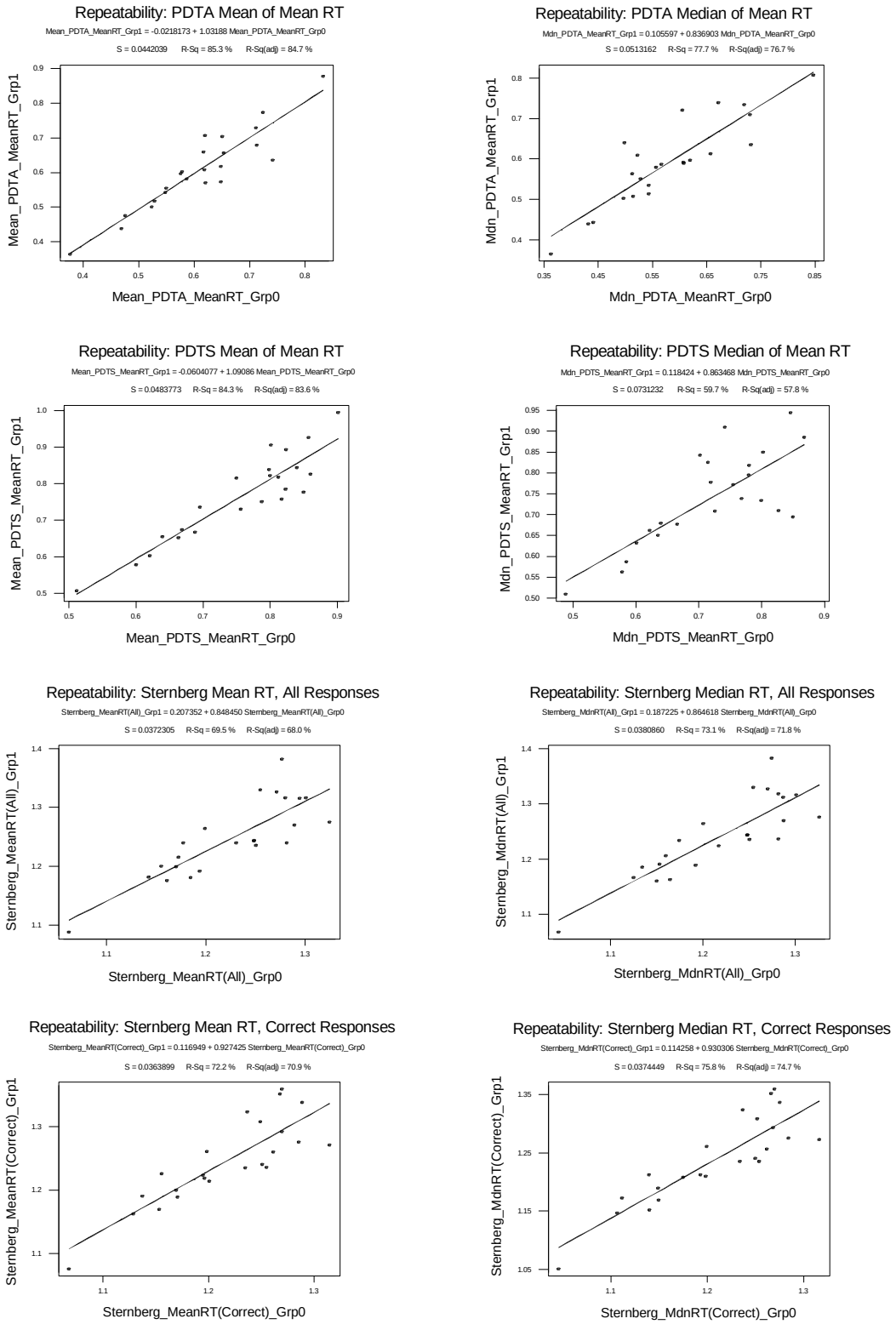


Figure 5-6. OED Detection: Group 0 Versus Group 1 Reaction Times for PDTA, PDTS, and Sternberg

5.1.1 Repeatability Summary

Repeatability results can be summarized as follows. Most of the surrogate measures were repeatable, defined as a correlation of about 0.70 or higher between Group 0 and Group 1 in the Split-Group analysis results. Scatter plots and regression lines provided additional insights. Moderate to high variability about the regression line was present in almost all measures. On the other hand, surrogate measures related to task time (Static Time, TSOT, and STISIM Task Duration) were exceptionally tight about the regression line. Task time values were for visual-manual tasks whose durations were task-intrinsic, i.e., inherent to the task itself. The length of a task was not arbitrarily set for these tasks. This result implies that task duration is largely limited or constrained by the nature of such tasks (number of steps, accuracy requirements, etc.) and is less subject to psychological effort or other sources of performance variation.

Mean values were better correlated than medians across Group 0 and Group 1 for several performance measures. This was indicated by a smaller standard error about the regression line than that found with medians. In other instances, there was no appreciable difference in the correspondence between mean and median values. Finally, Sternberg median Response Time measures showed slightly greater correspondence between Group 0 and Group 1 than with mean values. The mechanisms responsible for this pattern of effects are not known. It nonetheless suggests that different measures of central tendency may differ somewhat in terms of repeatability.

An interesting pattern of OED miss rate results emerged. As the participant became increasingly "loaded" by concurrent tasks, the repeatability of the measures improved. PDT Alone showed slightly more variability than PDT in STISIM. In turn, the Sternberg procedure produced even better correspondence. Increased repeatability appears to come with more demands placed on the performer that constrained performance.

There were a few unrepeatable (or insufficiently repeatable) surrogate measures. The Situational Awareness (SA) Magnitude Estimates scale was not repeatable. The SA ratings were collected once only due to scheduling limitations. Perhaps combining reps would have resulted in more stable results. Another possibility is that the judgments required of the test participants were hard to make. A few Sternberg measures were also not sufficiently repeatable. These measures included Reaction Time (RT) for Incorrect Responses and Percent Error given a detection.

5.2 Task Effects Results

Task effects charts were plotted only for repeatable surrogate measures. The charts provide a view of the typical values obtained for all tasks. Appendix Q contains an N-point summary (Hartwig and Dearing, 1979) that provides various summary statistics for each measure by task. What is not indicated on the charts or the N-point summaries is what pairs of tasks are statistically significantly different. This question will be addressed in the discriminability analysis section of this chapter.

The phrase Just Drive needs to be clarified when used in reference to test procedures that did not involve driving. This was done for convenience. In general, the Just Drive task in cases where there was no driving refers to the performance of that test procedure without a subsidiary task, e.g., Just Sternberg.

5.2.1 Operator Workload Scale and Multitasking Difficulty Magnitude Estimates

Subjective workload ratings of DWM task workload are presented together. Figure 5-7 presents task effects on OWL and Figure 5-8 presents Multitasking Difficulty scales for all DWM tasks. Average OWL scores ranged from a low of 30 to a high of 60. The mean OWL ratings of all of the higher-workload visual-manual tasks (as defined by prior prediction in Chapter 2) were greater than any of the lower-workload tasks. The lowest and highest of all mean OWL ratings were for visual-manual tasks. This suggests that visual-manual tasks may have been easier to rate with OWL because the set contained several conventional tasks and perhaps more obvious features of distraction potential. On the other hand, the average OWL ratings for auditory-vocal tasks were intermixed. Route Instructions was rated highest, on average, of the auditory-vocal tasks. But Book-on-Tape Summarize was rated higher than Travel Computations; Route Orientation was rated second from the bottom of the set.

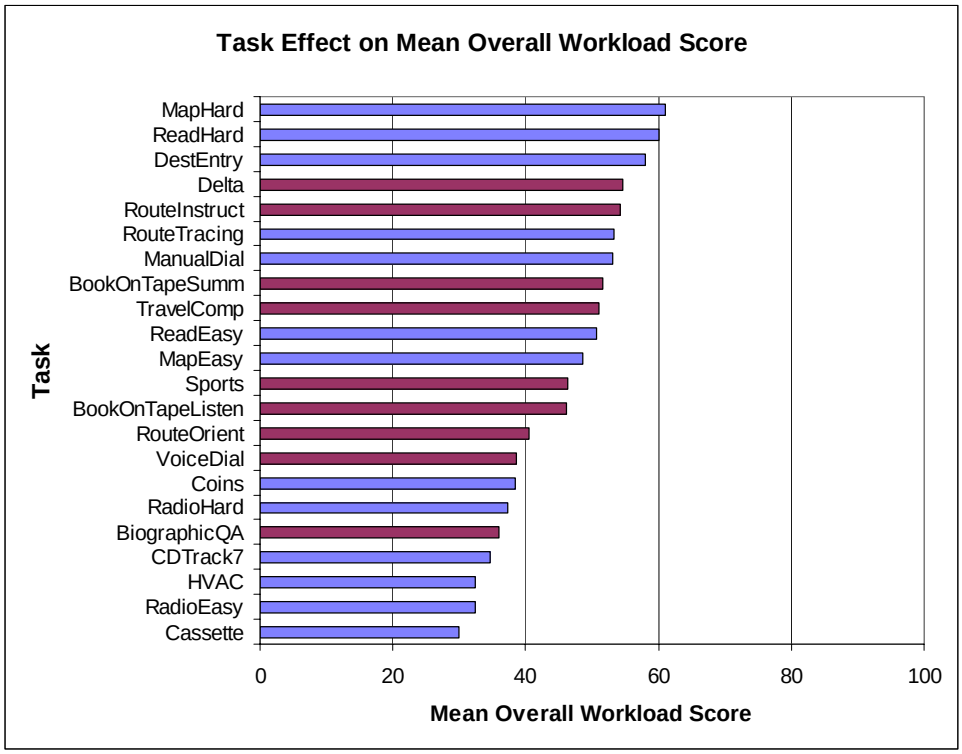


Figure 5-7. Task Effects on Operator Workload

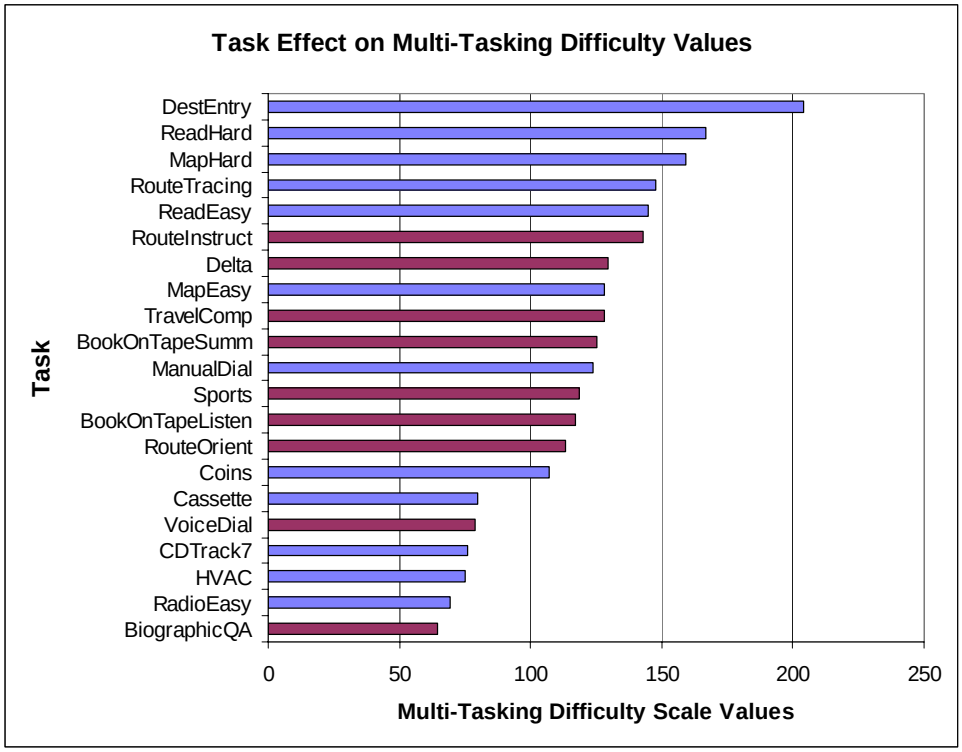


Figure 5-8. Multitasking Difficulty Magnitude Estimates

The Multitasking Difficulty magnitude estimates also ranked all higher-workload visual-manual tasks above those in the lower-workload category. See Figure 5-8. The auditory-vocal tasks were ranked somewhat better than with OWL with respect to prior prediction. Route Instructions and Travel Computations were rated highest on the MD scale. But Route Orientation was rated below Book-on-Tape Summarize, Sports Broadcast, and Book-on-Tape Listen. At the task level, the OWL and MD values were highly correlated ($r = 0.92$, $p < 0.001$, $df = 11$).

Either OWL or the Multitasking Difficulty ratings appeared to distinguish higher-workload from lower-workload tasks when applied to visual-manual tasks. This may reflect the more obvious effects of visual-manual task demands while driving such as eyeglances away from the road scene. The categorization performance of each of these subjective workload scales is less for auditory-vocal tasks. These types of tasks may be less obvious in their workload impacts. The two methods yield essentially the same information, though the MD scale reverses the Book-on-Tape Summarize and Travel Computations tasks in the right direction with respect to prior prediction.

The OWL and MD ratings were obtained by averaging two reps each. The ratings were made after instruction in the DWM tasks. The second rep was also taken after attempting to perform the DWM tasks in one or more laboratory test settings. Empirical data indicate that the OWL scale fares well as a subjective workload scale in a variety of settings (Hill et al., 1992). Unlike OWL's 100-point scale, magnitude estimation scales are unbounded ratios. The magnitude estimation method has a long history in psychology (Gescheider, 1997) and is one of the most widely used scaling methods. It is one of the few techniques that can, in theory, provide an absolute scale of the psychological dimension being assessed. The extent to which this was achieved merits further research and verification.

5.2.2 Static Time, Total Shutter Open Time, and the R-Metric

Typical Static Time, TSOT, and R-Metric values are presented in Figure 5-9, Figure 5-10, and Figure 5-11. The median Static Times across the tasks range from about 5 seconds (Radio (Easy)) to about 19 seconds (Read (Hard)). Destination Entry is exceptional in that it has a typical static completion time of almost 60 seconds. Median TSOT values show a similar ordering of tasks. Destination Entry again stands apart from the other tasks. Mean R-Metric ratios values vary from 0.59 (CD / Track 7) to 1.19 (Map (Hard)). The correlation between median Static Time and median TSOT is $r = 0.99$ ($p < 0.001$, $df = 11$). By contrast, the mean R-Metric has non-significant correlations (near zero) with these measures. This is an indication that the R-Metric assesses a different task attribute than TSOT and Static Time which both measure a task duration attribute. Mean R-Metric task effects are presented rather than the median values in deference to common use of mean values

TSOT is usually less than the Static Time for each task. This effect has been reported in other occlusion studies (Green and Tsimhoni, 2001). Purportedly, it arises because the occlusion shutter-closed period allows the test participant to continue working on at least some aspects of the task. Some of these aspects might include the following: complete a button press; page forward in a list; turn a knob approximately; plan for the next step in a task; etc.

Several different organizations have proposed criteria for these metrics to identify visual-manual tasks that should not be performed while driving. These criteria are often defined in terms of means or 85th percentile values. Table 5-5 presents both mean and 85th percentile values for Static Time, TSOT, and the R-Metric for each of the DWM tasks evaluated. Prior Predictions are also presented to aid interpretation. Various rules are provided to sort DWM tasks into higher-workload or lower-workload categories. Those categorizations that do not match Prior Prediction are indicated with an asterisk.

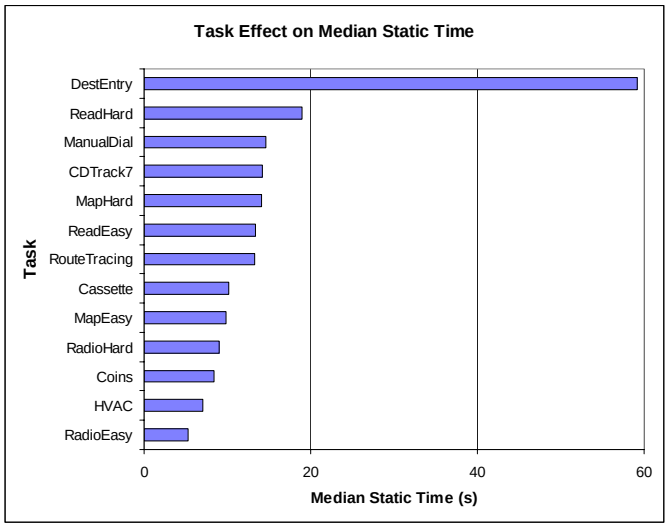


Figure 5-9. DWM Visual-Manual Task Effects for Static Time

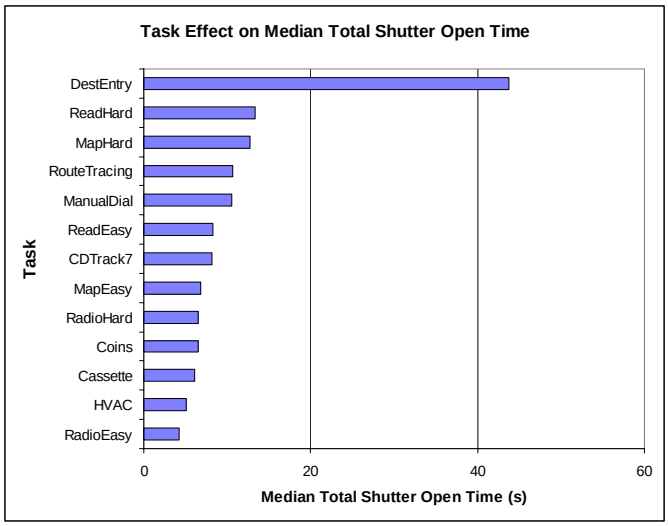


Figure 5-10. DWM Visual-Manual Task Effects for Total Shutter Open Time

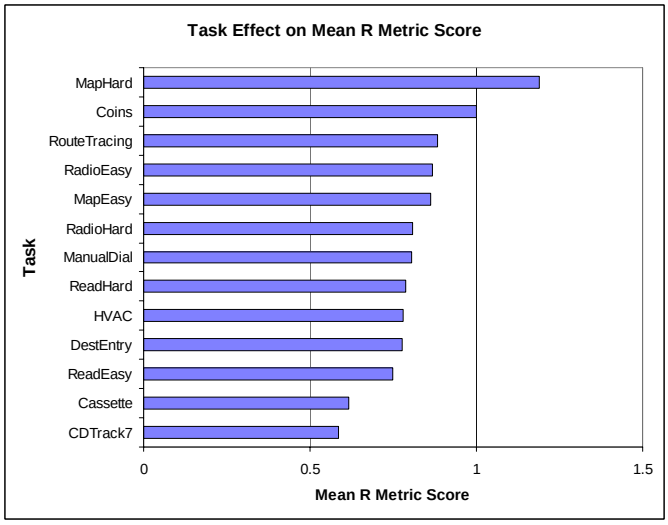


Figure 5-11. DWM Visual-Manual Task Effects for R-Metric

Table 5-5. Various Higher-Workload Versus Lower-Workload Visual-Manual Task Sorting Rules: Static Time, TSOT, and R-Metric Values

Task Type	Prior Prediction	Rule 1 A ≥ 15 s	Rule 2 B ≥ 15 s	Rule 3 C > 15 s	Rule 4 D > 15 s	Rule 5 C > 7.5 s	Rule 6 D > 7.5 s	Rule 7 F > 1.0	A Mean Static Time	B 85th %-ile Static Time	C Mean TSOT	D 85th%-ile TSOT	E Mean R-Metric	F 85th %-ile R-Metric
Radio(Easy)	Lower	Lower	Lower	Lower	Lower	Lower	Lower	Higher*	5.97	7	4.66	6.6	0.87	1.14
HVAC	Lower	Lower	Lower	Lower	Lower	Lower	Lower	Lower	8.00	9.9	5.55	7.4	0.78	0.99
Cassette	Lower	Lower	Higher*	Lower	Lower	Lower	Higher*	Lower	11.15	15.8	6.37	8	0.62	0.8
Coins	Lower	Lower	Lower	Lower	Lower	Lower	Higher*	Higher*	9.68	14.8	6.98	9.1	1.00	1.5
Radio(Hard)	Lower	Lower	Lower	Lower	Lower	Lower	Higher*	Lower	9.70	11.7	7.30	9.5	0.81	1
CD/Track7	Lower	Lower	Higher*	Lower	Lower	Higher*	Higher*	Lower	14.67	16.5	8.34	9.8	0.59	0.66
Map(Easy)	Higher	Lower*	Lower*	Lower*	Lower*	Higher	Higher	Lower*	10.43	13.3	8.06	10.6	0.86	1
Read(Easy)	Higher	Lower*	Higher	Lower*	Lower*	Higher	Higher	Lower*	14.23	18.4	9.59	12.1	0.75	0.96
Manual Dial	Higher	Higher	Higher	Lower*	Lower*	Higher	Higher	Lower*	16.65	23.2	11.66	14.9	0.81	0.99
Route Tracing	Higher	Lower*	Higher	Lower*	Lower*	Higher	Higher	Higher	14.14	17.25	11.55	15	0.88	1.13
Read(Hard)	Higher	Higher	Higher	Lower*	Higher	Higher	Higher	Lower*	19.26	24.4	13.22	15.8	0.79	1
Map(Hard)	Higher	Lower*	Higher	Lower*	Higher	Higher	Higher	Higher	14.62	20.8	13.41	18.3	1.19	1.29
Destination Entry	Higher	Higher	Higher	Higher	Higher	Higher	Higher	Lower*	64.14	79.1	48.53	62.8	0.78	0.95

Notes: Each Rule defines a higher-workload task
Each asterisk (*) indicates a difference with DWM Prior Expectation

A scan of Table 5-5 reveals that Rule 5 has the best classification performance. Rule 5 states that any task with a mean TSOT > 7.5 s indicates a higher-workload task. This rule reflects recent driver distraction research in Japan (Asoh et al., 2002). Other rules do not categorize as well. These results may support future discussions to set performance criteria with these surrogate metrics. Keep in mind that the detailed methods used in the DWM project may differ significantly from those proposed by others.

The R-Metric is intended to capture the ease with which a person may interrupt and resume a visual-manual task. If the R-Metric is above 1.0, then it is interpreted to indicate a relatively uninterruptible task. Only the Map (Hard) task has a mean R-Metric value above 1.0. This task may be more difficult to resume because it requires visual search to reorient to the task. Radio (Easy) and Coins tasks have 85th percentile values above 1.0, yet these are lower-workload tasks by prior prediction. It is not clear why these tasks would be relatively uninterruptible while comparable short tasks (HVAC), or comparable heavily manual tasks (e.g., Insert Cassette) would not show similar results. Route Tracing, and Map (Hard) tasks also have 85th percentile R-Metric values above 1. Both of these tasks may be relatively uninterruptible because they require visual search to reorient to the task once the driver looks away.

5.2.3 STISIM Measures

Figure 5-12 shows DWM task effects on STISIM Task Duration, STISIM Standard Deviation of Lane Position, STISIM Percent Lanex Cross trials, and STISIM Speed Difference (SpeedDiff). STISIM Task duration shows auditory-vocal tasks separated from visual-manual tasks (except for Destination Entry). This reflects the fact that auditory-vocal tasks (except the summary task) were set to an approximately two-minute duration each. Differences among the auditory-vocal task durations reflect variations in length of recorded materials; participant task completion (e.g., some may have responded even before a simple biographic question was finished); and experimenter variability when marking the task duration start and end points. On the other hand, the visual-manual task durations are inherent to the task. They are not arbitrary, given the goal that the task a participant was given.

The SDLP results show that lanekeeping variation was greatest for the visual-manual tasks, generally in agreement with prior prediction. The auditory-vocal tasks, through longer in duration, generally had lower typical SDLP values. Auditory-vocal tasks allowed for more looking toward the road scene. Very short duration tasks such and HVAC and Radio (Easy) were below Just Drive in terms of median values. It is also worth noting that some auditory-vocal tasks of comparable duration had lower SDLP values than Just Drive with both mean and median values. This suggests that some concurrent task load improved focus to what was otherwise a mundane driving task.

SpeedDiff showed a greater influence of task duration then SDLP. Longer auditory-vocal tasks generally had the greatest Speed Diff values. However, Destination Entry and Route Tracing stand out among the visual-manual tasks in typical SpeedDiff values. These results held for both mean and median SpeedDiff values.

Percent Lanex (Cross) trials showed two large task effects: Destination Entry and Route Tracing. Other visual-manual task effects are smaller but largely consistent with prior prediction. In contrast, the auditory-vocal tasks tend to have lower lanex events associated with them.

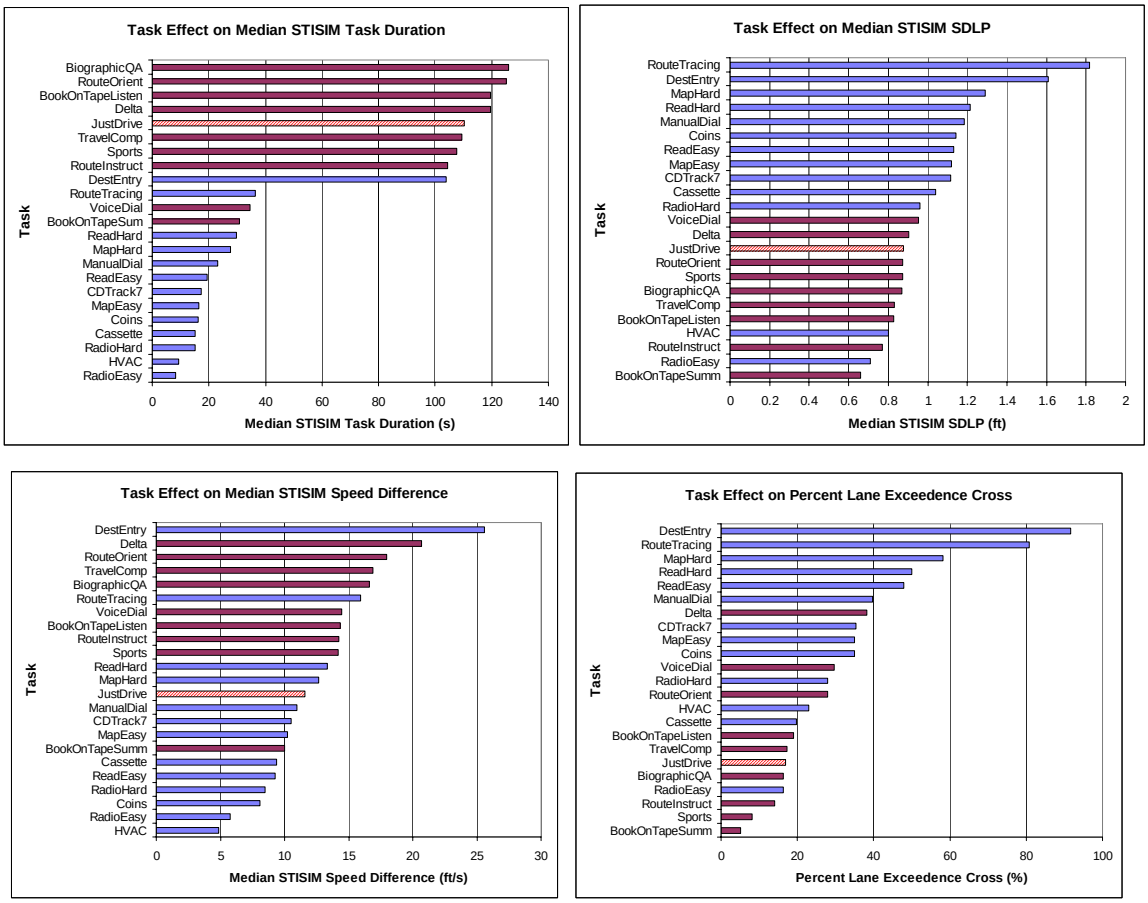


Figure 5-12. Typical DWM Task Effects for Selected STISIM Measures

Figure 5-13 presents a draftsman's plot (Chambers, Cleveland, Kleiner, and Tukey, 1983) of the selected STISIM surrogate driving measures. This figure is essentially a visual correlation matrix among the variables. The variable name at the top of each column defines the variables plotted in each cell. For example, the upper-left scatter plot presents the median STISIM duration (the x variable) and mean STISIM SDLP (the y variable). As another example, consider the middle scatter plot on the bottom row. It is a plot of median STISIM SDLP (the x variable) and Percent Lanex (Cross) Trials (the y variable). Note the strong linear relationships between mean and median SDLP and between mean and median SpeedDiff measures. This is to be expected since both are measures of central tendency for the same sample of data. Note also the high correlation between SDLP and Percent Lanex (Cross) trials. Such a relationship suggests that, at least within STISIM, serious lapses in lanekeeping (i.e., lane line crossings) are a continuation of growing lane position variation. The other scatter plots show two distinct sets of tasks largely broken out by task duration. The shorter visual-manual tasks generally show a linear relationship between median STSIM Task Duration and other surrogate measures. Task duration, therefore, appears to play a substantive role in at least some surrogate measures of driving.

Draftsman's Plot Of Selected STISIM Surrogate Measures

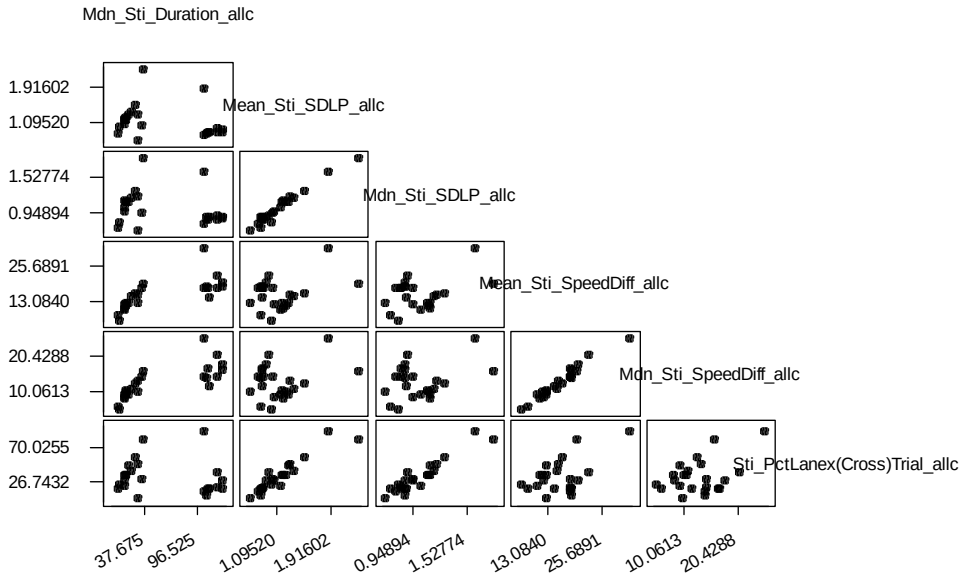


Figure 5-13. Draftsman's Plot of Selected STISIM Surrogate Measures

5.2.4 OED Measures: PDT and Sternberg Miss Rates and Reaction Times

Object and Event Detection (OED) is an important part of driving. It is the first event necessary for subsequent decision and response. A variety of methods were tested in the laboratory to examine OED performance during DWM task performance. The PDT-Alone (PDTA) method required only a simple button press to indicate a detection of the light stimulus. The PDT in STISIM (PDTS) method required the same simple button press as PDTA, but now the DWM task was done concurrently with driving. The Sternberg method required more cognitive complexity. The test participant memorized three road signs before each trial. He or she then had to choose between a "Yes" response if a Sternberg probe presented while doing a task was one of the memorized signs or a "No" response otherwise. This method provided, in principle, an assessment of working memory load and elementary (yes or no) decision processes associated with event detection.

Figure 5-14 presents Miss Rate measures for each of the methods. Notice the wider range of miss rates with PDTS as compared with PDTA. This increased spread of miss rates across tasks is presumably due to the added load of concurrent driving in the simulator. The visual-manual tasks are mostly higher than any of the auditory-vocal tasks with the PDT methods. This reflects the visual nature of the PDT stimuli as well as the manual demands of driving, concurrent task performance, and the button press response. The Insert Cassette task may be exceptional because it has a heavy manual component (Wierwille, 1993). The mixed-mode tasks lay between the visual-manual and auditory-vocal tasks with both PDT methods. The mixed-mode tasks have both visual-manual as well as auditory-vocal components. Finally, notice that HVAC has a higher PDTS missed detection rate than Destination Entry. Manual Dial has a higher PDTA missed detection rate than Destination Entry.

Sternberg miss rates showed a somewhat different picture. No distinction is made between Sternberg Spatial and Sternberg Verbal results because these differences were found to be not significant. Sternberg Percent Missed Detections are similar to the PDT miss rates because they are all measures of no response. The magnitudes of Sternberg missed detections are, overall, somewhat greater than that for the PDT methods. This may be due to differences in how conspicuous the displays were, though both PDT lights and Sternberg LCD displays were well above threshold. It may also reflect greater effort in making the Yes or No decision that would delay a response to the point where it was declared a missed detection.

The range of Sternberg task effects is relatively smaller than for the PDT data. Except for a shift of the Book-on-Tape Summarize task, the visual-manual tasks are again associated with more missed detections than the auditory-vocal tasks. Again, the mixed-mode tasks lay between these other two task sets. And again, HVAC is associated with worse performance than Destination Entry. The Combined Decrement score shifts several auditory-vocal tasks up to the high end of decrement. This may reflect, at least in part, the multiple-components that are present in these tasks. For example, each of these auditory-vocal tasks contains intermittent checks on performance, any one of which might be scored incorrect even though the task itself is rated as partially successful. Visual-manual tasks, even those with multiple subtasks, are generally either successful (if the goal state is reached) or not (if the goal state is not reached). There is no evaluation for intermediate errors that are corrected, such as overshoots or undershoots in radio tuning, problems inserting a cassette, correcting a wrong number in a manual dial, and so on.

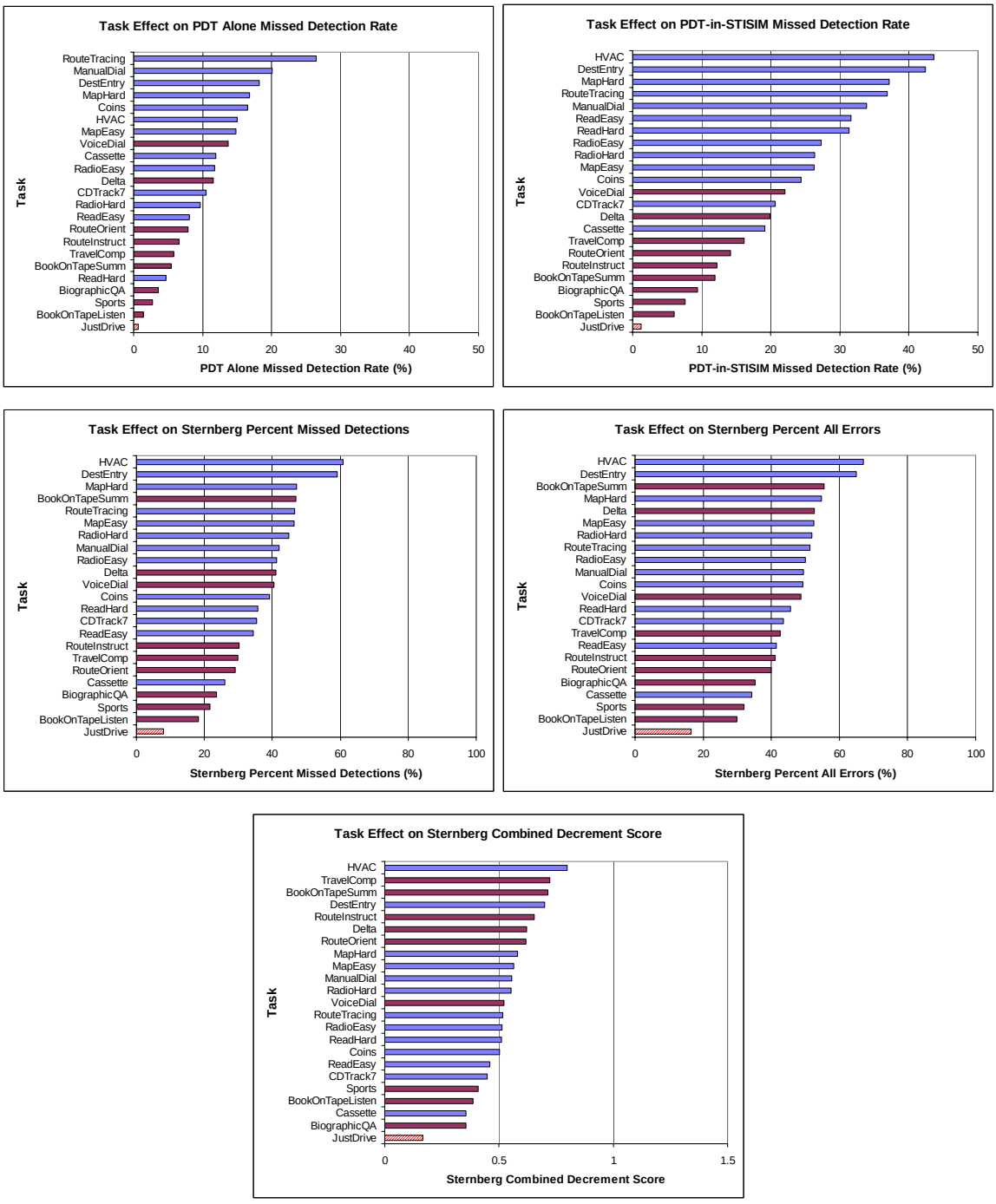


Figure 5-14. Object-and-Event Detection Miss Rates for Various Laboratory Measures

Sternberg Reaction Time data for both all responses and for correct (yes or no) responses are given in Figure 5-15. The most apparent feature of these figures is the separation between visual-manual and auditory-vocal tasks. The longer RTs for visual-manual tasks probably reflect the conflict between manual resource demands mentioned earlier. The second feature of these figures is the narrow range of typical RTs.

PDTA and PDTS mean and median response time results are also presented in Figure 5-15. PDTA mean and median results differ primarily in the shift of the Destination Entry task in the rank order of effects. A difference between mean and median values might have been due to the skewness of a distribution. Otherwise, the results are consistent regardless of what measure is used.

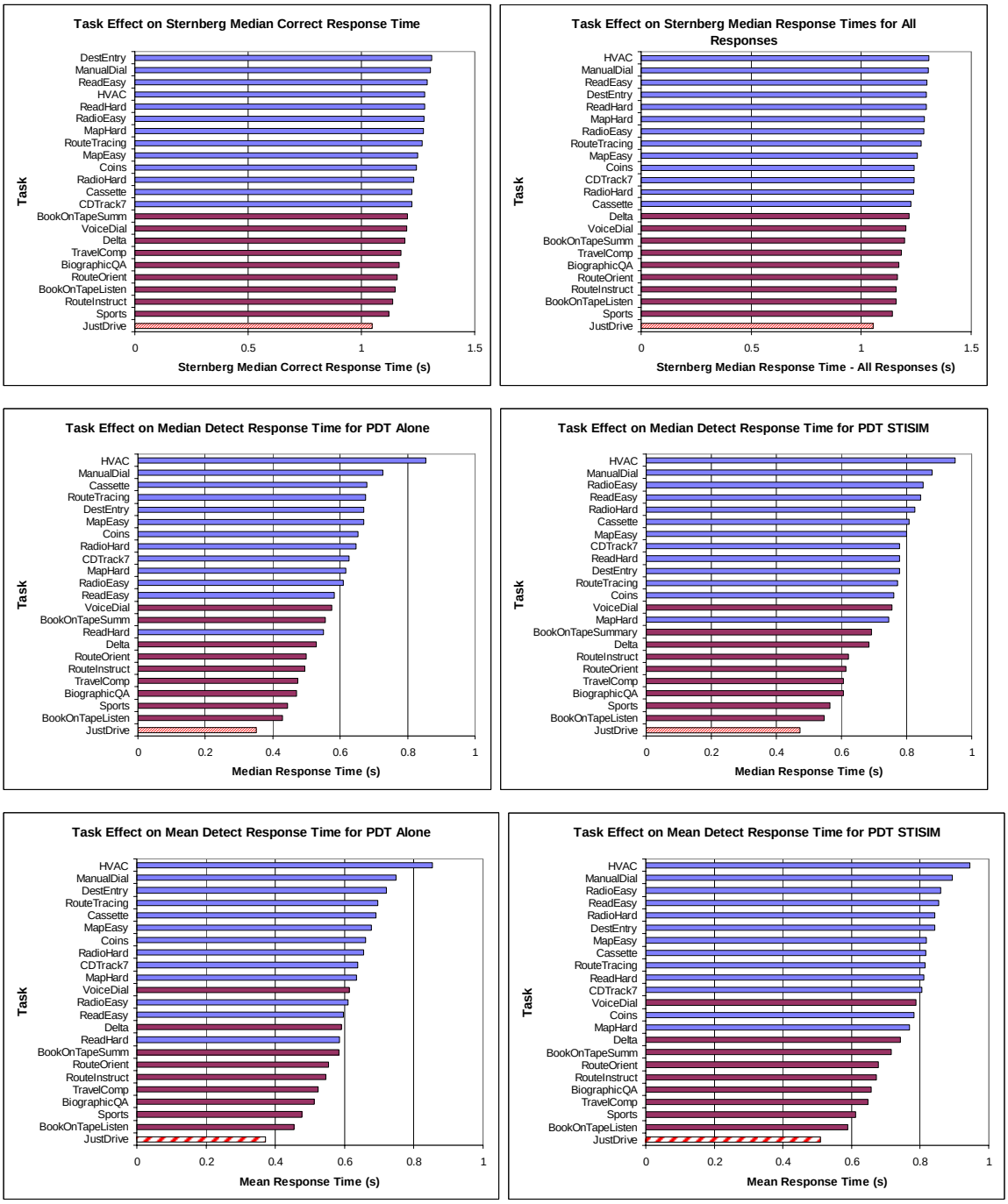


Figure 5-15. Sternberg and PDT Response Times

5.2.5 Task Effects Summary

A summary of task effects will be presented after the discriminability results are presented.

5.3 Discriminability Analysis

This analysis evaluated the sensitivity of a surrogate measure to discriminate between higher-workload versus lower-workload tasks. There are (23*22)/2 or 253 pairwise comparisons possible for 23 tasks. A more manageable number of comparisons were needed. This was achieved by categorizing tasks into higher-workload or lower-workload categories based on the literature, modeling, and content analysis. (See Chapter 2, *Study Design Overview*, for details of the prior predictions). Visual-manual tasks and auditory-vocal tasks were separated into two sets. The prior predictions approach reduced the number of visual-manual task comparisons to 42. The prior predictions approach reduced the number of auditory-vocal (plus Just Drive) comparisons to 15. The Just Drive task was grouped with the auditory-vocal tasks because, like them, Just Drive had been arbitrarily set to last approximately two minutes in duration. (Book-on-Tape Summarize was the sole auditory-vocal exception to the fixed duration). The visual-manual tasks, on the other hand, had task durations that were intrinsic to the task. Like Auditory-vocal tasks, Just Drive would not involve any additional visual input or manual output task demands. Finally, distraction or inattention effects would likely center on working memory for both auditory-vocal and Just Drive tasks. Mixed-mode tasks (Voice Dial and Flight Information (Delta Flightline)) were not assessed in the discriminability analysis because these tasks are less well understood.

Paired comparisons were made using directional tests for added power. The non-parametric sign test was used to avoid data truncation or data transforms needed for parametric procedures (Conover, 1999; Marascuilo and McSweeney, 1977). Generally speaking, the paired comparisons tested the likelihood that participants had higher workload scores for a given higher-workload task than the scores for a given lower-workload task that they also performed. A per comparison alpha level of 0.05 was used throughout.

No procedures were used to control experiment-wise error, i.e., the overall likelihood of at least one statistically significant individual test outcome obtained by pure chance. The procedures to manage this problem reduce the power of individual tests, often severely. Mature areas of science may properly reduce the risk of a false-positive result by increasing the risk of a false negative result. The DWM project was largely exploratory research. As such, it seemed better to uncover potentially useful findings for further investigation or application than to prematurely or inadvertently throw away such findings. The reader should nonetheless interpret the results with caution.

Discriminability values for each selected surrogate measure were prepared in the following way. The selected laboratory surrogate measures presented here generally increase as workload increases. So, the sign test evaluated the experimental hypothesis that participant measures were greater for a higher-workload task than for a lower-workload task. Statistically significant differences were tallied. The total number of statistically significant tests divided by the total number of tests was operationally defined as the Percent Significant, offered as an index of discrimination power for that surrogate measure. Only comparisons of tasks across categories were made. This managed the number of pairwise comparisons to be made. Note that the same approach is applicable to measures for which lower values may be associated with higher workload (e.g., mirror glances decrease, duration of glimpses to the road scene during a visual-manual task shorten, etc.). The direction of the one-tail test is simply reversed. Multiple comparisons were made for visual-manual tasks separate from auditory-vocal tasks.

As described in Chapter 3, two different levels of surrogate discriminability were examined. The first level was discriminability of a task effect from Just Drive. A subsidiary task carried out

while driving would likely have some effect distinct from just driving alone. Therefore, it was considered important to examine a measure's ability to distinguish between subsidiary task effects from Just Drive (or performance on the primary laboratory surrogate task alone). The second level of discriminability compared higher-workload to lower-workload tasks to distinguish tasks across the two categories.

The sign test is carried out at a level below a measure's means or medians or percents. The sign test takes the difference between each participant's scores on two different tasks and ignores the magnitudes of these differences. If there really is no difference between the two tasks, then only random variation will contribute any positive and negative differences. If a workload measure is thought to increase as workload increases, a higher-workload task would have a greater value on that measure than a lower-workload task. This would result in more positive differences than negative differences. The sign test then assesses the likelihood that the observed proportion of positive differences is significantly greater than 50-50. If a workload measure is thought to decrease as workload increases, then the sign test assesses the likelihood that the observed proportion of negative differences is significantly greater than 50-50. This is why the discriminability results do not refer to means, medians, and the like.

The discriminability results are summarized in Table 5-6. Discriminability percentages greater than or equal to 70 percent are highlighted. While somewhat arbitrary, measures with discriminability values below 70 percent might be considered insufficiently sensitive for application.

Table 5-6. Laboratory Discriminability Results for Selected Surrogate Measures

Laboratory Surrogates Discriminability Summary

Based on Sign Test

	Auditory-Vocal Tasks				Visual-Manual Tasks			
Metric	**Column 1: Low Workload In-Vehicle Tasks vs. Just Drive**	**Column 2: High Workload In-VehicleTasks vs. Just Drive**	**Column 3: All In-Vehicle Tasks vs. Just Drive**	**Column 4: High Workload Tasks vs. Low Workload Tasks**	**Column 5: Low Workload In-Vehicle Tasks vs. Just Drive**	**Column 6: High Workload In-VehicleTasks vs. Just Drive**	**Column 7: All In-Vehicle Tasks vs. Just Drive**	**Column 8: Low Workload Tasks vs. High Workload Tasks**
OWL Rating	Note 1	Note 1	Note 1	67%	Note 1	Note 1	Note 1	95%
Multitasking Difficulty	Note 1	Note 1	Note 1	60%	Note 1	Note 1	Note 1	95%
Static Time	Note 2	Note 2	Note2	Note 2	Note 2	Note 2	Note 2	81%
TSOT	Note 2	Note 2	Note 2	Note 2	Note 2	Note 2	Note 2	88%
R-Metric	Note 2	Note 2	Note2	Note 2	Note 2	Note 2	Note 2	38%
STISIM Duration	50%	33%	43%	40%	0%	0%	0%	86%
STISIM SDLP	0%	0%	0%	27%	67%	100%	85%	69%
STISIM %Lanex Trials	0%	33%	14%	33%	33%	100%	69%	79%
STISIM SpeedDiff	75%	100%	86%	40%	0%	57%	31%	79%
PDTa Miss Rate	75%	100%	86%	93%	83%	86%	85%	43%
PDTs Miss Rate	100%	100%	100%	93%	100%	100%	100%	62%
StrnPctMiss	100%	100%	100%	67%	100%	100%	100%	36%
StrnPctAllError	100%	100%	100%	73%	100%	100%	100%	36%
StrnCombDecr	100%	100%	100%	80%	100%	100%	100%	43%
StrnMeanRTAll	100%	100%	100%	33%	100%	100%	100%	26%
StrnMeanRTCorr	100%	100%	100%	27%	100%	100%	100%	17%
PDTA Mean RT	100%	100%	100%	73%	100%	100%	100%	24%
PDTS Mean RT	100%	100%	100%	67%	100%	100%	100%	5%

Notes

Note 1: Discriminability scores are not available for the the OWL or MultiTasking Difficulty metrics. These surrogates were not run with the Just Drive task.

Note 2: The Occlusion and Static Time surrogates were not run with the Auditory-Vocal Tasks or with Just Drive. No discriminability scores are available for these metrics.

Note 3: Discriminability scores of 67% or higher are highlighted.

Note 4: For Column 4, Lower Workload Auditory-Vocal Tasks included Just Drive. See text for explanation.

Note 5: For Column 8, Higher Workload vs. Lower Workload Task comparisons for Visual-Manual Tasks did not include Just Drive. See text for explanation.

Note 6: Number of comparisons : 4 for Column 1; 3 for Column 2; 7 for Column 3; 15 for Column 4; 6 for Column 5; 7 for Column 6; 13 for Column 7; 42 for Column 8

5.3.1 Discriminability Results: Auditory-Vocal Tasks and Just Drive

Table 5-6, left-hand side, presents the paired comparison results for the auditory-vocal study tasks and Just Drive. The first three columns present the percentage of statistically significant paired comparisons of task results from Just Drive (or Just Surrogate) test results. The fourth column presents the percentage of higher-workload versus lower-workload task comparisons that were statistically significant. The number of comparisons behind the percentages was often very small. There were four lower-workload tasks compared against Just Drive in Column 1. Three higher workload tasks were compared against Just Drive for Column 2; there were seven comparisons for Column 3. There were 15 comparisons for Column 4, three higher-workload tasks compared to five lower-workload tasks (including Just Drive as a lower-workload task).

Consider the results for the two repeatable subjective assessment scores. Just Drive was not evaluated with these methods (see Note 1 of Table 5-6). Thus, there were $k = 12$ paired comparisons for three higher-workload auditory-vocal tasks compared to four lower-workload auditory-vocal tasks. (See Chapter 2 for prior predictions). Based on these comparisons, subjective assessments did not adequately distinguish among higher- versus lower-workload tasks at the 70 percent level. However, OWL was marginally acceptable.

Static Time, Total Shutter Open Time (TSOT), and R-Metric values were not collected for auditory-vocal or Just Drive tasks. Static time was not appropriate for the DWM auditory-vocal tasks because these tasks were set to approximately fixed durations. TSOT and the R-Metric were applicable only for visual-manual tasks.

An interesting pattern of results was obtained with STISIM driving performance surrogate measures. Only SpeedDiff was able to adequately distinguish auditory-vocal tasks from Just Drive. The STISIM Task Duration results were driven by the shorter duration of the Book-on-Tape Summarize task and minor but consistent predetermined task durations. None of the driving performance measures was an adequate discriminator of higher-workload from lower-workload tasks.

Various OED measures showed a markedly different pattern of results. All the OED measures were able to distinguish between the auditory-vocal tasks and Just Drive, usually with 100 percent discrimination. Almost all OED miss rate data also distinguished well between higher- and lower-workload tasks. It should also be noted that OED reaction time measures were less discriminating than the miss rate data obtained with the same method.

5.3.1.1 Summary of Auditory-Vocal and Just Drive Discriminability Results

Subjective assessments were not adequately discriminating among the DWM auditory-vocal tasks. Static time, TSOT, and R-Metric measures could not be applied to these tasks. Driving performance measures were not good discriminators for this class of DWM tasks. However, OED measures were good discriminators among auditory-vocal and Just Drive tasks. The difference between vehicle measures and OED results may be explained by the selective withdrawal of attention (Tijerina, 2000; Brown, 1994). Eye glances away from the road scene reflect a general withdrawal of attention from the road. The visual demands of a task can influence both vehicle control and visual OED performance. Auditory-vocal tasks do not require eyeglances away from the road scene. This allows highly practiced driving skills like lanekeeping to remain largely unaffected. On the other hand, OED performance may be selectively affected by cognitive load. SpeedDiff, though able to distinguish tasks from Just Drive, nonetheless could not distinguish well between tasks of categorically different workload levels.

5.3.2 Discriminability Results: Visual-Manual Tasks

Table 5-6, right-hand side, presents the paired comparison results for the DWM visual-manual tasks and Just Drive. The first three columns present the percentage of statistically significant paired comparisons of task results from Just Drive (or Just Surrogate) results. The fourth column presents the percentage of higher-workload versus lower-workload task comparisons that were statistically significant. The number of comparisons behind the percentages were often very small. There were six lower-workload tasks compared against Just Drive in Column 5. Seven higher-workload tasks were compared against Just Drive for Column 6. There were 13 comparisons for Column 7. There were 42 comparisons for Column 8, seven higher-workload tasks compared to six lower-workload tasks (Just Drive was not included as a lower workload task due to its considerably longer and fixed duration).

Subjective assessments did not include Just Drive so associated comparisons were not appropriate. Thus, only Column 8 has percentages for the two repeatable subjective measures. As can be seen, OWL was an excellent discriminator between higher-workload and lower-workload visual-manual tasks. However, results indicated that participants did not sufficiently discriminate Manual Dial from HVAC or Read (Easy) from Coins. Multitasking Difficulty magnitude estimation was also an excellent discriminator. Participants did not discriminate Manual Dial from HVAC (again) discriminate or Manual Dial from Coins with this procedure. Such excellent discriminability may arise because of the more apparent nature of visual-manual task demands.

Static Time and Total Shutter Open Time (TSOT) measures had excellent discriminability. The TSOT measure was somewhat more sensitive than Static Time in distinguishing between higher-workload and lower-workload tasks. The R-Metric, did not distinguish among the DWM visual-manual tasks. This may be because the R-Metric did not adequately capture interruptibility or because the DWM tasks were largely interruptible.

STISIM driving performance surrogate measures discriminated among the DWM visual-manual tasks in several ways. STISIM Task Duration showed no significant differences against Just Drive. This is to be expected given the directional test predicted that greater workload would be associated with longer task durations. The visual-manual tasks were uniformly shorter than the 2 minute Just Drive task, including the Destination Entry task.

STISIM Task Duration was sensitive to task differences between higher-workload and lower-workload tasks. This is an important differentiation because task duration was intrinsic to the visual-manual tasks rather than arbitrarily fixed in duration. SpeedDiff did not adequately distinguish either lower-workload or higher-workload tasks from Just Drive. Given the typically short durations of visual-manual tasks, there was less opportunity for speed to vary much. On the other hand, SpeedDiff did an acceptable job in discriminating higher-workload tasks from lower-workload tasks.

Standard Deviation of Lane Position was only marginally discriminable. It did discriminate higher-workload tasks from Just Drive. But it did not discriminate lower-workload tasks from Just Drive. Furthermore, it was not quite acceptable to differentiate between visual-manual tasks (without the 2-minute Just Drive included) in terms of prior workload prediction.

A discrete measure of lanekeeping, Lanex (Cross) trials did discriminate well between higher-workload tasks and Just Drive. It did a better job of discriminating between higher-workload and lower-workload visual-manual tasks. This may indicate that different processes underlay SDLP (a continuous lanekeeping measure) and Lane Exceedances (a discrete measure of a breakdown in lanekeeping). Another possibility of differences in the results is a by-product of the discrete versus continuous nature of the measurements.

OED measures present a very different pattern of results as compared to the auditory-vocal tasks and Just Drive. Like the auditory-vocal tasks, OED measures overall distinguished visual-manual tasks from Just Drive. These measures were uniformly poor in discriminating among visual-manual tasks in terms of prior prediction. This result may be driven largely by the paradoxical nature of the OED data for visual-manual tasks. Again, OED response time measures were less discriminable than miss rates obtained with the same methods.

5.3.2.1 Summary of Visual-Manual Task Discriminability Results

Both OWL and Multitasking Difficulty magnitude estimates were excellent at discriminating between higher-workload and lower-workload visual-manual DWM tasks. This stands in contrast to the poorer discriminability results with auditory-vocal tasks.

TSOT did as well or better than any other surrogate performance metric to differentiate between visual-manual tasks based on prior prediction. Static Time was also good, though not as good in discriminability. Finally, the R-Metric did not discriminate between lower-workload and higher-workload tasks. This latter result may be because most tasks were interruptible.

STISIM driving performance measures had good discriminability overall. Task Duration, SpeedDiff, and Lanex (Cross) trials were almost as good as TSOT in discriminating between higher-workload and lower-workload visual-manual tasks. SDLP was marginally acceptable with a discrimination value of 69 percent versus the stated criterion of 70 percent. OED discriminability was uniformly poor between visual-manual tasks with respect to prior prediction. This is due to the paradoxical nature of these OED results, as indicated in the Task Effects charts presented earlier. The OED results were, however, able to distinguish these tasks relative to Just Drive.

This pattern of results illustrates several points. A measure may discriminate among tasks from different categories yet be poorly related to driving performance. OWL, for example, may be an excellent indicator for customer perceptions of task difficulty, but not of driving performance. A measure may discriminate among tasks from different categories, yet have only very specific relationships to driving performance. Sternberg Percent Missed Detections, for example, is only related to Percent Follow Vehicle Turn Signal (FVTS) Miss Rate.

Some surrogate methods may be unfairly assessed with this comparison procedure. For example, the R-Metric is intended to provide an index of the ease with which tasks may be interrupted and resumed later. This dimension of task demand is not captured in the prior predictions developed for this project. Furthermore, the DWM task may simply not be sufficiently varied in interruptibility to allow for a better assessment of this metric.

There is also legitimate concern that the tasks are sometimes not as different in demand as the DWM task designers had hoped. For example, Map (Easy) is often not significantly different from Coins, Insert Cassette, or CD / Track 7 across various surrogates. These tasks may, in fact, have greater similarities than differences, e.g., in the spare capacity of the participant to monitor the other concurrent task. Different aspects of the tasks (and task similarity in those aspects) are likely to be assessed to different degrees by different surrogate measures.

TSOT appears to be as good or better surrogate metric than others evaluated for visual-manual tasks. While this method has been available for some time, the DWM project has contributed more empirical data, collected over a broader set of tasks, and with a different participant sample than past or concurrent projects. It should be kept in mind, though, that different surrogates may be differentially sensitive to different kinds of tasks. That is, even if the percentage of significant differences is low, those few comparisons that were significant may be useful in their own right. The discriminability results are global indicators and ignore such fine distinctions.

5.4 Prediction of Selected Driving Performance and Eyeglance Measures with Laboratory Surrogates

Surrogate predictions were assessed separately for track data and road data. Repeatable surrogate measures were correlated with various driving performance and eyeglance measures. Scatter plots were then prepared for measures correlated at $r \geq 0.70$. Scatter plots provided greater insights into the nature of the correlations.

Results are presented for track measures and then for road measures. Within each venue, visual-manual task results are presented first. Then auditory-vocal, Just Drive, and mixed-mode task results are presented. The rationale for grouping the latter tasks together will be presented later.

Laboratory participants were scheduled to complete each task with each of the surrogates. A different sample of participants performed each completed task on the Track. A third group of participants, different from the other two, completed road trials. Three separate groups of participants were used to include more variability among the participants. Correlations between surrogates and driving measures would be maximized if the same persons performed in all venues. In practice, customer or marketing clinics do not work like that. A participant sample in a clinic is intended to provide data that might be predictive of how a different, larger group of customers might react.

Laboratory surrogates were not driving measures or eyeglance measures. They were intended to be proxies for driving measures or eyeglance measures. For example, OWL scale values (in dimensionless units) might be correlated to SDLP (in units of ft/s). Even STISIM trials should not be considered equivalent to driving on the road or on the track. Surrogates and road or track measures could not be correlated at the test participant level because different test participants were used in each test venue. Thus, correlations were made at the DWM task level of summary statistics like medians or percents. As a result, correlations and regressions reported here are based on small numbers of cases:

- Track trials
  - 13 visual-manual tasks (sometimes Destination Entry was omitted and this reduced the task set to 12 visual-manual tasks)
  - 10 auditory-vocal, mixed-mode, and Just Drive tasks
- Road trials
  - 7 visual-manual tasks (HVAC, Radio (Easy), Radio (Hard), Coins, CD/Track 7, Insert Cassette, and Manual Dial)
  - 9 auditory-vocal, mixed-mode, and Just Drive tasks (Delta Flight Information was not attempted on the road)

Small numbers of cases require higher levels of correlation to be statistically significantly different from zero. A lower limit of $r \geq |0.70|$ was set as a significant correlation for two reasons. This value was based on the lower limit of critical values for statistical significance for the small numbers of cases. Also, a squared value of r of approximately 0.50 is interpreted as the proportion of variance in the driving or eyeglance measure that co-varies with, or is accounted for by, the laboratory surrogate measure. This is often used as a lower limit for applied human factors work (e.g., Kroemer, 1986).

The results will be presented in the four sets of findings used in the Task Effects section. Subjective assessments (OWL and Multitasking Difficulty scales) are analyzed together. Static Time, TSOT, and the R-Metric are examined as a set of related measures. STISIM driving performance measures are analyzed separately. Finally, surrogate OED measures are treated together.

With specific exceptions (e.g., OWL, Multitasking Difficulty scales), medians were used instead of means for prediction even when both measures were repeatable. Medians are preferred as typical values because they are more resistant than means to extreme values or skewness in the data.

Prediction results are presented first for visual-manual tasks performed on the test track. Only the track venue allowed visual-manual tasks of all potential distraction levels to be evaluated. Thus, it represents the most complete set of visual-manual task effects.

Auditory-vocal tasks and Just Drive results from Track trials are presented subsequently. For the prediction analyses, the auditory-vocal and Just Drive tasks were augmented to also include the mixed-mode tasks of Voice Dial and Delta Flight Information. These tasks were added for the following reasons. First, the mixed-mode tasks were considered to be primarily auditory-vocal in nature. The test participant picked up and dialed manually but then continued through the task by working with an interactive voice response system. Second, the inclusion of the mixed-mode tasks increased the number or tasks available with which to carry out the regression and correlation analyses. Note that the mixed-mode tasks were not included in the Discriminability results. This omission was because less is known about these tasks with respect to prior prediction. In particular, the performance characteristics of an interactive voice response system can make the ranking of these tasks into higher-workload and lower-workload categories unclear.

5.4.1 Prediction of Selected Track Driving Performance and Eyeglance Measures with Laboratory Surrogates: Visual-Manual Tasks

The task effects of the Destination Entry (DestEntry) task were often extreme in the surrogate measures and track measures. Therefore, this task was removed from these analyses to provide more sensible rather than misleading correlation and regression results. The Destination Entry task served as a high leverage point if it was outside the trend of the other tasks and needed to be removed. In such cases, removal was justified because that task was atypical of the remainder of the data. On the other hand, if the Destination Entry task was within the trend of the remaining data but an extreme point nonetheless, the remaining tasks would adequately reflect that trend anyway. If Destination Entry was removed in this case, the benefit would be in reducing an artificially inflates R-square value. In either case, the Destination Entry task effects were not subtle. Correlations between selected repeatable surrogate measures and selected track measures are presented in Table 5-7.

The table suggests that there was a reasonably high correlation between each track measure and at least one surrogate measure. Scatter plots and draftsman's plots, shown later, indicate that this was not necessarily so. Some track measures are not well predicted by any DWM surrogates. These track measures include Center High-Mounted Stoplight (CHMSL) detection performance; task-related mean single-glance times; the proportion of time spent on task-related glances away from the road scene; task-related glances per second.

Table 5-7. Laboratory Surrogates and Track Correlations: Visual-Manual Tasks Without Destination Entry Task

Lab vs.Track Correlations:
Visual-Manual Tasks (no DestEntry Task included)

Measure	Task Duration	Lateral Control		Longitudinal Control	Event Detection			Task-Related Selected Eye Glance Measures				
Lab vs. Track	Median Task Duration	Median SDLP	%Lanex Cross Trials	Median SpeedDiff	%LVD MissRate	%CHMSL MissRate	%FVTS MissRate	Mean Glance Duration	Mean No. Of Glances	Mean Total Duration Of All Glances	Proportion Of Task Time	Mean Glance Rate
Mean OWL	0.86	0.70	-	0.87	-	-	-	-	0.85	0.85	-	-
Multi-tasking Difficulty Scale	0.86	0.75	-	0.85	-	-	-	-	0.82	0.82	-	-
Mean Static Time	0.91	-	-	0.89	-0.83	-	-	-	0.94	0.86	-	-
Median Static Time	0.91	-	-	0.88	-0.80	-	-	-	0.94	0.88	-	-
Mean TSOT	0.95	0.74	-	0.95	-0.73	-	-	-	0.94	0.92	-	-
Median TSOT	0.95	0.69	-	0.92	-0.73	-	-	-	0.95	0.93	-	-
Mean R Metric	-	-	-	-	-	-	-	-	-	-	-	-
STISIM Mean Task Duration	0.96	0.87	-	0.92	-	-	-	-	0.89	0.88	-	-
STISIM Median Task Duration	0.96	0.89	-	0.92	-	-	-	-	0.88	0.87	-	-
STISIM Mean SDLP	0.72	0.91	0.82	0.75	-	-	-	-	-	-	-	-
STISIM Median SDLP	0.81	0.97	0.76	0.83	-	-	-	-	-	-	-	-
STISIM Percent Lanex Cross	0.85	0.92	-	0.83	-	-	-	-	0.79	0.81	-	-
STISIM Mean Speed Difference	0.91	0.92	-	0.92	-0.72	-	-	-	0.85	0.84	-	-
STISIM Median Speed Difference	0.91	0.89	-	0.93	-0.71	-	-	-	0.82	0.79	-	-
PDT Alone Percent Miss Rate	-	-	0.73	-	-	-	-	-	-	-	-	-
PDT with STISIM Percent Miss Rate	-	-	-	-	-	-	-	-	-	-	-	-
Sternberg Percent Missed Detections	-	-	-	-	-	-	0.75	-	-	-	-	-
Sternberg Percent All Errors	-	-	-	-	-	-	0.82	-	-	-	-	-
Sternberg Combined Decrement	-	-	-	-	-	-	0.82	-	-	-	-	-

- denotes correlations with p > 0.05 or r < 0.70

A number of draftsman's plots were prepared to gain further insight into the prediction relationships. The draftsman's plot, as indicated earlier, is a visual correlation table that shows the relationships between pairs of variables. It is read by considering the variable label at the top of a column to be the x-variable for that column. The remaining variable labels to the side are each a y-variable that is correlated with or predicted by the x-variable. Only significant correlations were represented in the Draftsman's plots.

Consider the subjective workload assessments. Figure 5-16 presents a draftsman's plot of the OWL and Multitasking Difficulty subjective assessments (Columns 1 and 2) that were correlated with their associated Task-Related (TR) driving performance and eyeglance measures on the Track (Rows 3 through 6). The OWL and MD scales are loosely correlated with the track measures. None of the trends is particularly well structured. All exhibit considerable scatter, except those for Task Duration, SpeedDiff, Mean Glances TR and MeanduratTR (or total glance time to task).

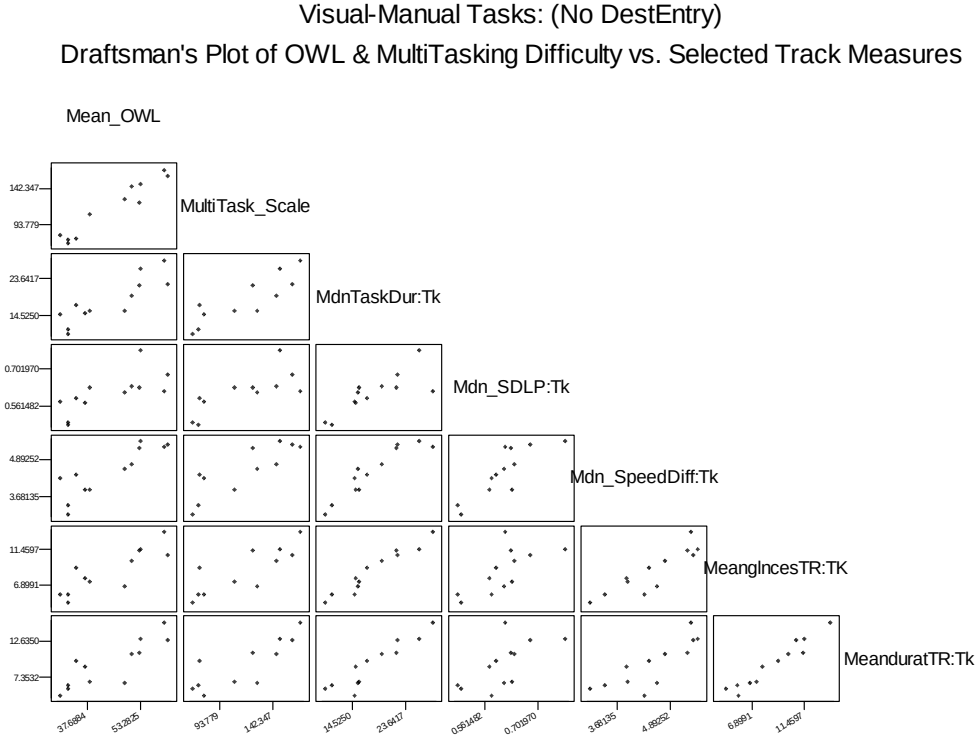


Figure 5-16. OWL, Multitasking Difficulty Scale, and Selected Track Measures (No Destination Entry Task)

Static Time and TSOT were considered together for two reasons. First, they are both appropriate for time-intrinsic visual-manual tasks. Second, they both contribute to the R-Metric, an estimate of the ease with which a task can be resumed after an interruption. Figure 5-17 presents a draftsman's plot of Static Time and TSOT surrogate measures along with selected track driving performance and eyeglance measures. The TSOT measure appears to be as good or better a surrogate predictor than Static Time. TSOT is strongly correlated with Task Duration and average Task-Related glance counts. Median SpeedDiff and Task-Related total glance duration away from the road are also highly correlated but with more variability. Finally, the negative correlation

between TSOT and lead vehicle deceleration (LVD) is very diffuse. It results mainly because of two short tasks (HVAC and Radio (Easy)) and by perceptual factors discussed elsewhere in this report. Figure 5-18 presents plots of TSOT versus Task Related mean glance counts both with and without Destination Entry. The extreme point follows the same trend as the other tasks.

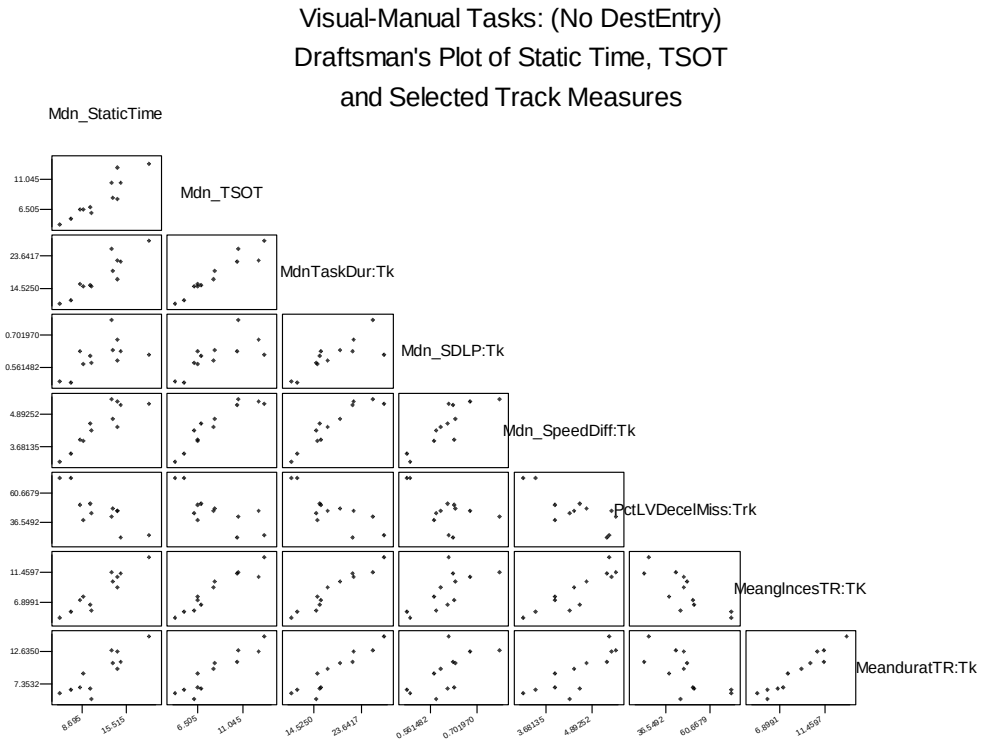


Figure 5-17. Draftsman's Plot of Static Time and TSOT Surrogate Metrics With Selected Track Metrics

STISIM driving performance measures were also compared to track driving performance measures. Figure 5-19 contains a draftsman's plot of STSIM measures and significantly correlated track driving performance measures. The plot indicates strong correlations between STISIM surrogate measures and their associated track measures. The exception is Percent Lanex (Cross) trials, which was not predicted by any of the STISIM measures. Route Tracing was associated with the highest percentages of lane exceeds (besides Destination Entry) in both DWM laboratory and track trials. This task is the extreme point at the top of all the plots with Percent Lanex (Cross) trials (Row 6). Finally, the negative correlation between the percent missed LVD detections and STSIM SpeedDiff are due to primarily to the presence of two short duration tasks in the upper left-hand corner of the associated graph (HVAC and Radio (Easy)).

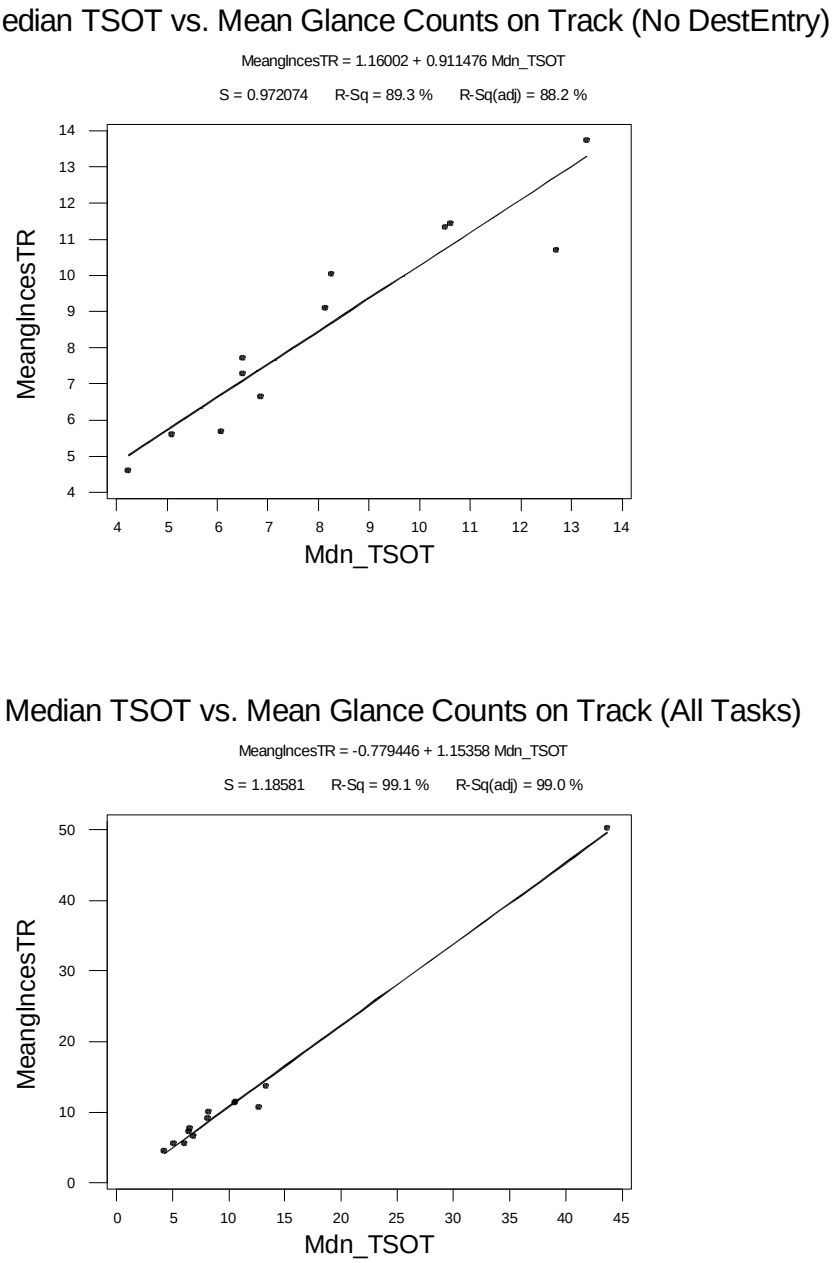


Figure 5-18. TSOT Versus Task Related Mean Glance Counts With and Without Destination Entry Task

The final group of laboratory results for the visual-manual tasks on the test track is OED surrogates. Figure 5-20 presents Draftsman's plots of Sternberg OED measures and Track Miss rates for the Follow Vehicle Turn Signal stimulus. The plot indicates strong correlations among the Sternberg measures themselves. However, the correlations between Sternberg measures and the track FVTS percent miss data are largely due to two extreme points. The HVAC task is the extreme point in the upper right-hand corner of the associated graphs (Row 3). There appears to be no linear structure with the plotted points between these extremes.

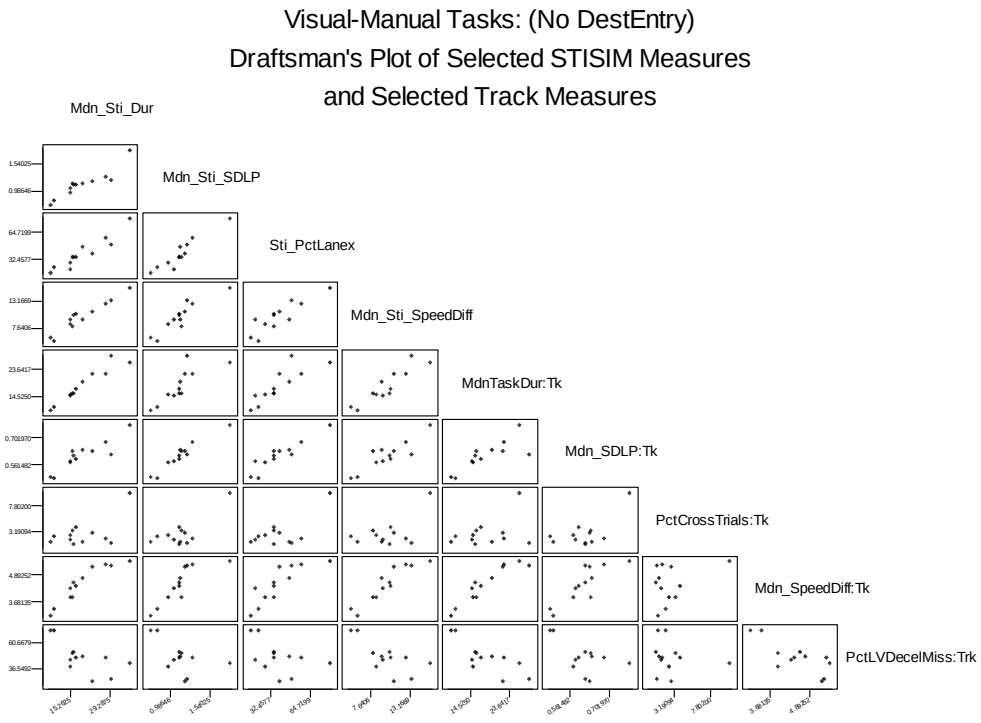


Figure 5-19. Draftsman's Plot of STISIM Metrics and Selected Track Metrics

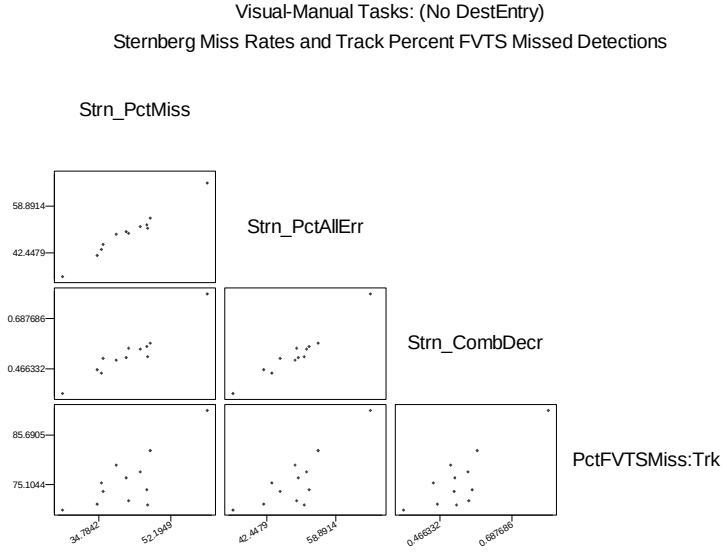


Figure 5-20. Draftsman's Plot of Sternberg OED Measures and Track FVTS Miss Rates

5.4.1.1 Summary of Laboratory Surrogate Prediction with Track Visual-Manual Tasks

The previous data presentations provide the following picture for prediction of track measures for visual-manual tasks.

- OWL and the MD subjective workload were correlated with selected track driving and eyeglance measures.
- Static Time and TSOT were both good predictors of eyeglance counts and total eyes-off-road time. Both were good predictors of Task Duration on the track, as well as SpeedDiff. TSOT was slightly better correlated than Static Time. The R-Metric was not correlated with any track measures.
- STISIM measures were very good predictors of comparable track driving measures. These measures included Task Duration, SDLP, and SpeedDiff. On the other hand, Percent Lanex Cross Trials on the test track was not predicted by STISIM lane exceeds. STISIM Percent Lanex Cross Trials served as a good predictive measure of SDLP on the Track.
- No Track OED percent missed detections were predicted by any of the laboratory surrogates.

5.4.2 Prediction of Selected Track Driving Performance Measures With Laboratory Surrogates: Auditory-Vocal, Mixed-Mode, and Just Drive Tasks

Table 5-8 presents the correlations between various laboratory surrogates the Track measures for the auditory-vocal, mixed-mode, and Just Drive tasks. The logic of including the latter in this set has been given earlier. Similarly, a rationale has been given for why mixed-mode tasks were not included in the auditory-vocal and Just Drive discriminability analysis but the Just Drive task was included.

Task-related eyeglance measures have been removed from this table for auditory-vocal tasks. However, see Chapter 3, *Test Track* Results and Chapter 4, *On-Road Results* for more in-depth analyses and discussion of eyeglance behavior in different task settings.

There was a high correlation between STISIM median task duration and track median task duration. This high correlation was largely due to the fact that the bulk of these tasks were of fixed duration. Remaining sources of duration variation would be due to variation in the task-intrinsic duration of the mixed-mode tasks and the Book on Tape Summarize task. Note also that no track driving performance measures were adequately predicted by any of the surrogate measures.

Track OED performance has several potential surrogate predictors. Figure 5-21 presents OWL and MD workload scale values along with track Percent CHMSL Missed Detections and Percent FVTS Missed Detections. Mean OWL was loosely positively correlated with the track FVTS miss rates. There was a negative correlation between the Multitasking Difficulty scale and track CHMSL miss rates. The draftsman's plot indicates that this is an artifact of the two tasks in the upper left-hand corner of the corresponding graph (Row 2, Column (2). Without these tasks (Voice Dial and Biographical Q&A), there appears to be no linear structure.

Table 5-8. Laboratory Versus Track Correlation for Auditory-Vocal, Mixed-Mode, and Just Drive Tasks

Lab vs. Track Correlations:
Auditory Vocal, Mixed Mode and Just Drive Tasks

Measure	Task Duration	Lateral Control		Longitudinal Control	Event Detection		
Lab vs Trk	Mdn Task Duration	Mdn SDLP	%Lanex Cross Trials	Mdn SpeedDiff	%LVD MissRate	%CHMSL MissRate	%FVTS MissRate
Mean_OWL	-	-	*	-	-	-	0.781
MultiTaskDiffScale	-	-	*	-	-	-0.728	-
Mdn_Static Time	*	*	*	*	*	*	*
Mdn_Sti_SDLP	-	-	*	-	-	-	-
Sti_%Lanex Cross	-	-	*	-	-	-	-
Mdn_Sti_SpeedDiff	-	-	*	-	-	-	-
Pdta_%MissRate	-	-	*	-	-	-	-
Pdts_%MissRate	-	-	*	-	-	-	-
Mdn_Sti_Tsk Dur	0.996	0.898	*	-	-	-	-
Strn_Pct missed detections	-	-	*	-	-	-	0.814
Strn_Pct all errors	-	-	*	-	-	-	0.853
Strn_Mdn Correct RT	-	-	*	-	-	-	-
Strn_Mdn All RT	-	-	*	-	-	-	0.756
Strn_CombDecr	-	-	*	-	-	-	0.88

- denotes correlations with p > 0.05 or < 0.70

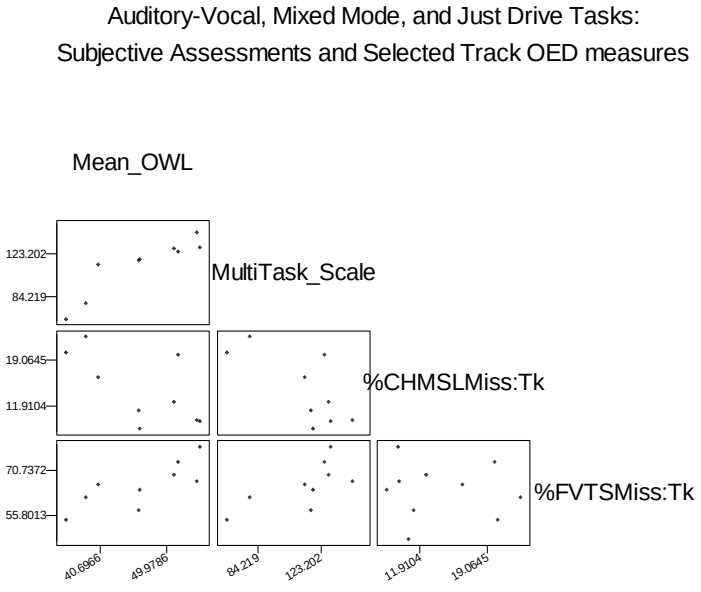


Figure 5-21. Draftsman's Plot of Subjective Workload Scales and Selected Track OED Measures for Auditory-Vocal, Mixed-Mode, and Just Drive Tasks

Several Sternberg measures appeared to be correlated with track FVTS miss rates. Figure 5-22 presents scatter plots of these measures. The Sternberg Combined Decrement score appeared to be the best predictor of FVTS miss rates. Figure 5-23 presents a scatter plot with each plotted point labeled by the task's name. There was a strong linear trend among most of the tasks. Just Drive had the lowest FVTS miss rate. Book-on-Tape Summarize (a short task) had the highest FVTS miss rate except for the Delta Flight Information task. The Delta Flight Information task and Sports Broadcast task were not in line with the other tasks.

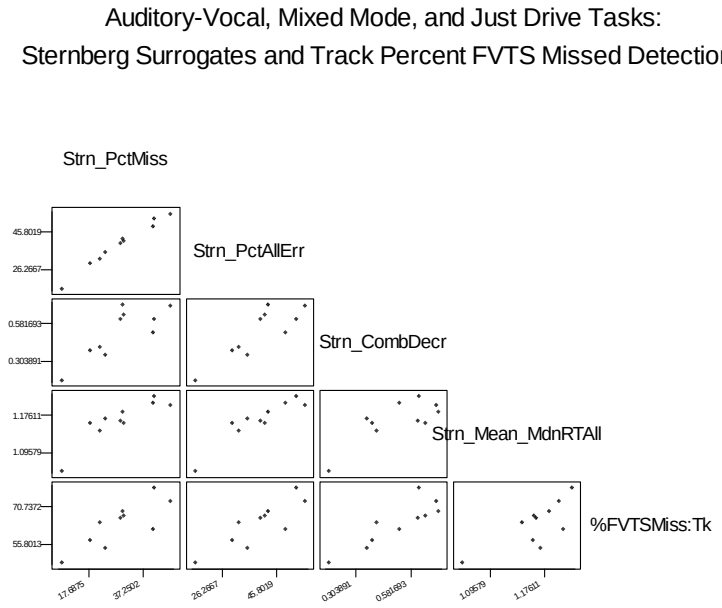


Figure 5-22. Draftsman's Plot of Sternberg Measures and Track Percent FVTS Missed

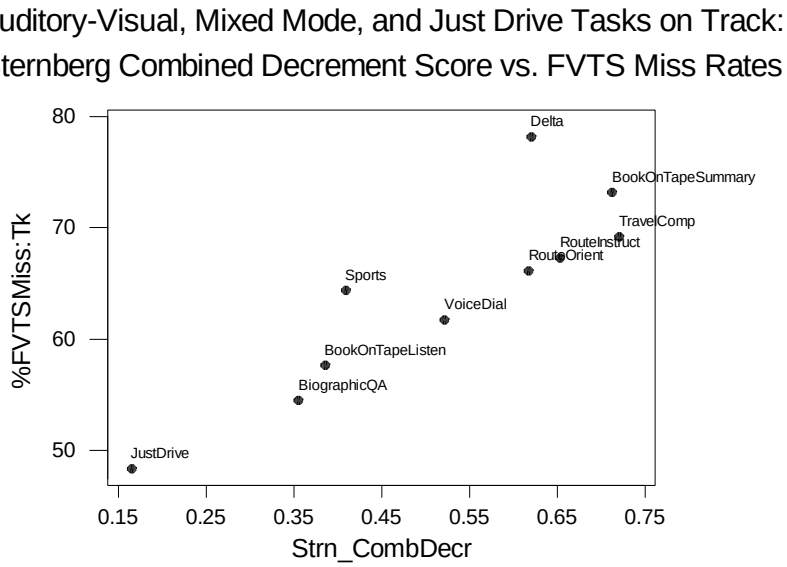


Figure 5-23. Sternberg Combined Decrement Score Versus FVTS Miss Rates

5.4.2.1 Summary of Laboratory Surrogate Prediction With Track Auditory-Vocal, Mixed-Mode, and Just Drive Tasks

The auditory-vocal, mixed-mode, and Just Drive tasks were markedly different than the visual-manual tasks in terms of surrogate prediction. Traditional task-related eyeglances measures were not applied due to the nature of these tasks. OWL and MD subjective workload scales were correlated with track OED measures either only loosely or as an artifact of one or more extreme points. No track driving measures were predicted by any of the selected surrogates. On the other hand, the Sternberg method produced useful predictions of track FVTS miss rates. In particular, the Combined Decrement Score provided very good prediction.

5.4.3 Prediction of Selected Road Driving Performance and Eyeglance Measures With Laboratory Surrogates: Visual-Manual Tasks

The visual-manual tasks tested on the road were a subset of the tasks tested on the track. Specifically, the road trials did not include visual-manual tasks that were thought to be more difficult. This had two effects on the outcomes of the road trials. There were only seven visual-manual tasks on the road rather than 13 on the track. With such a small number of cases, a correlation for statistical significance at the 0.05 level had to be greater than or equal to 0.67 (Edwards, 1984). This implies that as the number of data points grows smaller, the closer to a straight line those points must be to be significant. A second implication of the smaller number of tasks is that the range of visual-manual task effects was less than the range of effects found on the track.

Table 5-9 contains the correlations between various surrogate and road measures. There are no significant correlations for the LVD event. This is due, at least in part, to the short duration of these visual-manual tasks and the relatively long duration needed for the lead vehicle coast-down. There are also no significant correlations between any surrogate and either the task-related mean single-glance time or the proportion of task-related time away from the road ahead.

Table 5-9. Correlations for Laboratory Surrogates and Road Measures With Visual-Manual Tasks

Surrogate Correlations to On-Road Driving Performance Metrics for Visual Manual Tasks

Correlation Coefficent, r: Visual Manual Tasks											
Measures	**Task Duration**	**Lateral Control**		**Longitudinal Control**	**Event Detection**			**Task-Related Selected Eye Glance Metrics**			
Lab - Track	Median Task Duration	Median SDLP	Percent Lanex Cross Trials	Median Speed Difference	LVD MissRate	CHMSL MissRate	FVTS MissRate	Mean Glance Duration	Mean No. Of Glances	Mean Total Duration Of All Glances	Proportion Of Task Time
Mean OWL	0.910	--	0.810	0.777	--	--	--	--	0.980	0.913	--
Multi-tasking Difficulty Scale	0.907	0.824	0.962	0.874	--	--	--	--	0.900	--	--
Median Static Time	0.813	0.811	--	0.799	--	--	--	--	--	--	--
Median TSOT	0.956	0.861	0.820	0.895	--	--	--	--	0.818	0.755	--
Median R Metric	--	--	--	--	--	--	--	--	--	--	--
STISIM Median SDLP	0.837	0.937	0.874	0.911	--	--	--	--	--	--	--
STISIM Percent Lanex Cross	0.846	0.747	0.796	0.769	--	--	--	--	--	--	--
STISIM Median Speed Difference	0.793	0.855	--	0.847	--	--	--	--	--	--	--
PDT Alone Percent Miss Rate	--	--	0.754	--	--	--	--	--	--	--	--
PDT with STISIM Percent Miss Rate	--	--	--	--	--	0.806	0.810	--	--	--	--
STISIM Median Task Duration	0.837	0.937	0.874	0.911	--	--	--	--	0.777	--	--
Sternberg Percent Missed Detections	--	--	--	--	--	0.947	0.831	--	--	--	--
Sternberg Percent All Errors	--	--	--	--	--	0.969	0.862	--	--	--	--
Sternberg Median Correct RT	--	--	--	--	--	--	--	--	--	--	--
Sternberg Median All RT	--	--	--	--	--	--	--	--	--	--	--
Sternberg Combined Decrement	--	--	--	--	--	0.914	0.840	--	--	--	--

-- denotes correlations with $p > 0.05$

Figure 5-24 shows a draftsman's plot of OWL and MD Ratings (Columns 1 and 2) along with selected road measures. OWL predicts task-related glance counts (Column 1, Row 6) and Multitasking Difficulty ratings predict Road Percent Cross Trials (Column 2, Row 4). The remaining scatter plots contain a more diffuse distribution of points.

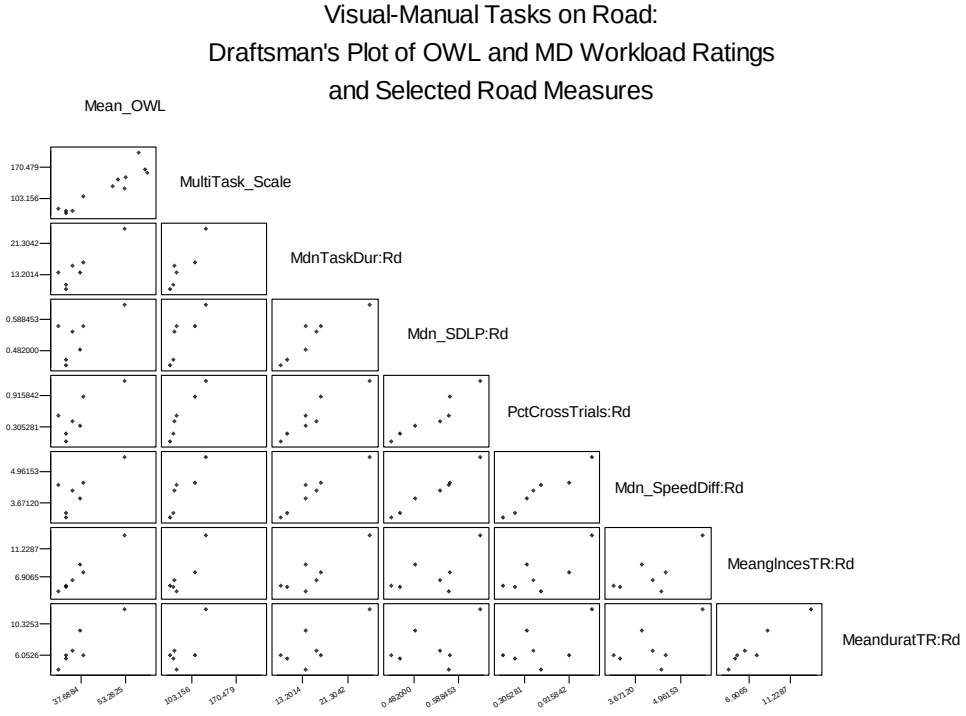


Figure 5-24. Visual-Manual Tasks On Road: Draftsman's Plot of OWL and Multitasking Difficulty Workload Scales

Median Static Time and TSOT were plotted against road measures with correlations above ± 0.70 (see Figure 5-25). The narrow range of these surrogates accounts for the nearly vertical scatter plots. The narrow range of task durations leads to similar very steep trend lines for STISIM Task Duration as a predictor for selected road measures (see Figure 5-26). Of the remaining surrogate measures, STISIM SDLP appears as a good predictor of several road measures (Column 2, Rows 5 through 7).

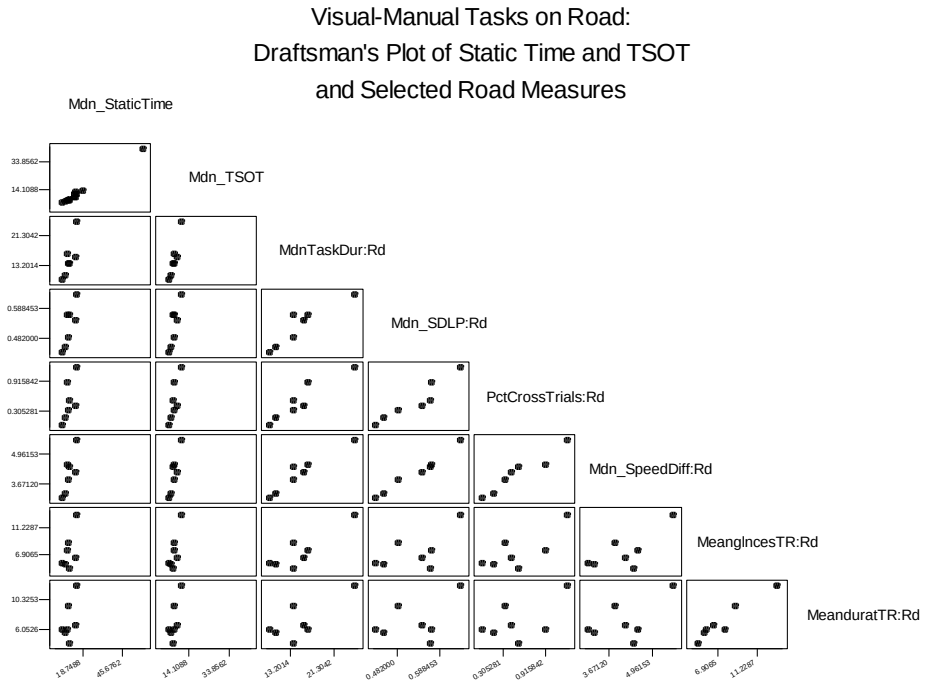


Figure 5-25. TSOT and Selected Measures for Visual-Manual Tasks

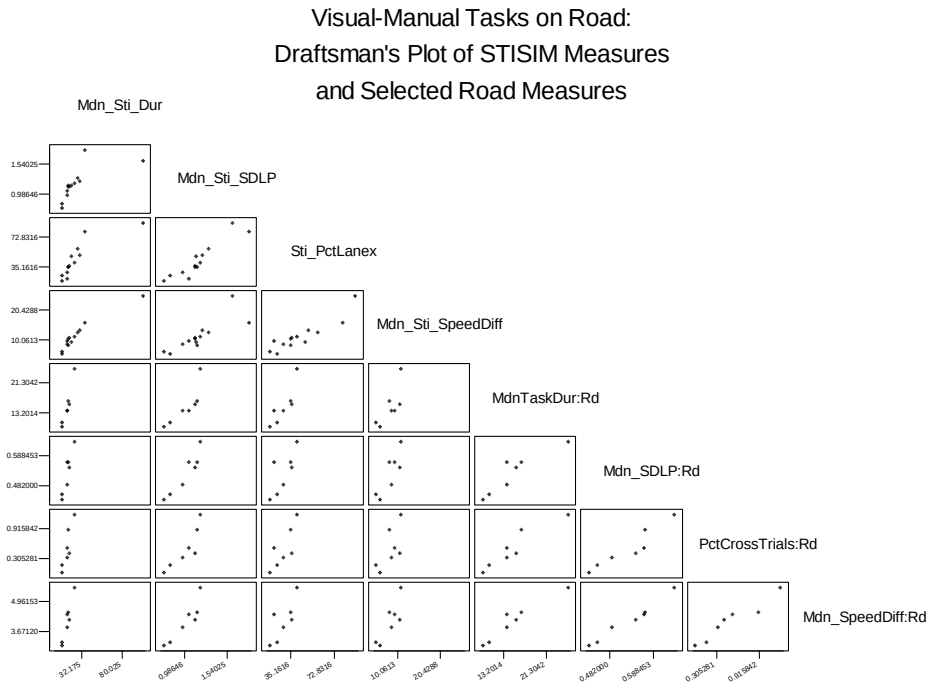


Figure 5-26. Selected STISIM Measures and Corresponding Road Driving Measures

Figure 5-27 contains scatter plots of various laboratory surrogate OED measures and road CHMSL miss rates (Row 7, Columns 3, 4, and (5) and FVTS miss rates (Row 8, Columns 3, 4,

and 5). The most structured trend is found with Sternberg measures and CHMSL miss rates. The other plots of surrogates and road OED measures do not show similar structure.

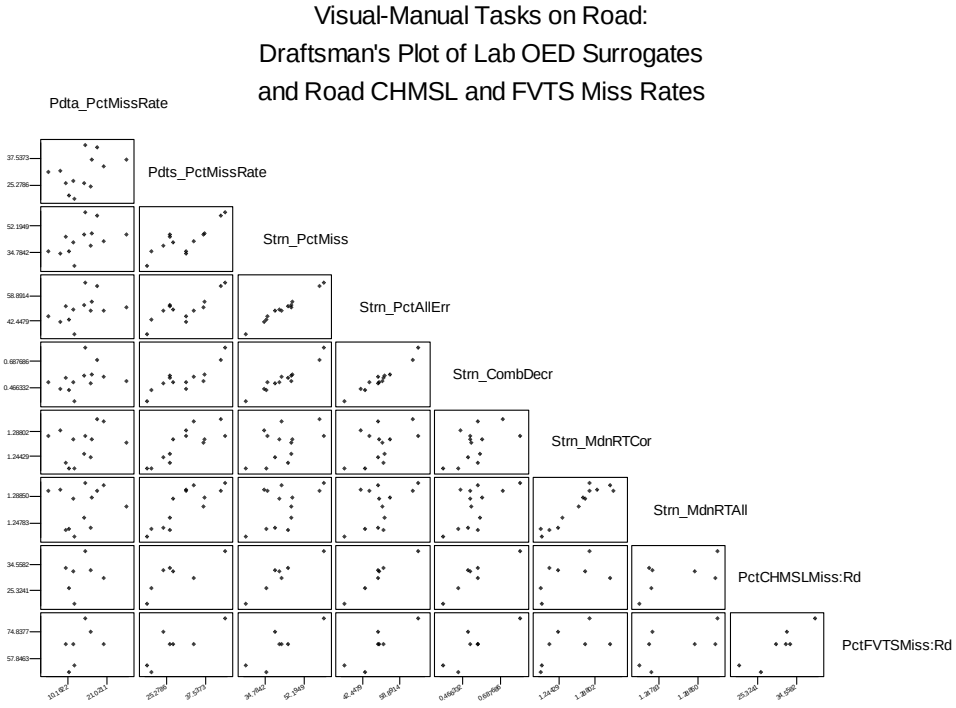


Figure 5-27. Draftsman's Plot of Laboratory OED Surrogates Versus On-Road CHMSL and FVTS Miss Rates for Visual-Manual Tasks

5.4.3.1 Summary of Laboratory Surrogate Prediction with Road Visual-Manual Tasks

The visual-manual tasks on the road were predicted by laboratory surrogates in various ways. The OWL and Multitasking Difficulty ratings were highly correlated with Task Related glance counts and Percent Lanex Cross Trials, respectively. Static Time and TSOT measures were distributed over a small range. This created nearly vertical trend lines with relevant road measures. STISIM measures had a similar problem. However, STISIM SDLP performed as well or better than other STISIM measure to predict road driving performance measures. Finally, several Sternberg measures were correlated with CHMSL and FVTS miss rates.

5.4.4 Prediction of Selected Road Driving Performance Measures With Laboratory Surrogates: Auditory-Vocal, Mixed-Mode, and Just Drive Tasks

The road trials did not include the Delta Flight Information task. Otherwise, the road trials used the same auditory-vocal, mixed-mode, and Just Drive Tasks as the track trials. Thus, nine tasks contributed to the results presented here.

Table 5-10 presents the prediction correlations for the road data. A number of correlations were not significant. Percent Lanex Cross and Percent LVD were not predicted by any surrogate. Subjective assessments did not predict any of the selected road measures. The STISIM Percent Lanex Cross Trials also did not predict any road measure.

Table 5-10. Prediction Correlations for Auditory-Vocal, Mixed-Mode, and Just Drive Tasks On-Road

Measure	Task Duration	Lateral Control		Longitudinal Control	Event Detection		
Lab vs. Road	Mdn Task Duration	Mdn SDLP	%Lanex Cross Trials	Mdn SpeedDiff	%LVD MissRate	%CHMSL MissRate	%FVTS MissRate
Mean_OWL	-	-	-	-	-	-	-
MultiTaskDiffScale	-	-	-	-	-	-	-
Mdn_Static Time	-	-	-	-	-	-	-
Mdn_Sti_SDLP	-	0.783	-	-	-	-	-
Sti_%Lanex Cross	-	-	-	-	-	-	-
Mdn_Sti_SpeedDiff	-	-	-	0.767	-	-	-
Pdta_%MissRate	-	-	-	-	-	0.836	-
Pdts_%MissRate	-	-	-	-	-	0.88	-
Mdn_Sti_Tsk Dur	0.998	-	-	0.793	-	-	-
Strn_Pct missed detections	-0.791	-	-	-	-	0.818	0.866
Strn_Pct all errors	-0.712	-	-	-	-	0.804	0.872
Strn_Mdn Correct RT	-	-	-	-	-	0.864	-
Strn_Mdn All RT	-	-	-	-	-	0.833	-
Strn_CombDecr	-	-	-	-	-	-	0.891

- denotes correlations with p > 0.05

STISIM and Sternberg surrogate correlations with road driving metrics are plotted in Figure 5-28. The plot of median STISIM task duration and median road task duration was highly linear with minimum scatter. This reflected the fact that the majority of these tasks were of fixed duration. Other plots between laboratory and road measures did not have clear structure. For example, the negative correlations associated with median SpeedDiff and Sternberg measures (Columns 4 and 5, Row 7) may be attributed to extreme values in the lower right-hand side of the graphs.

Auditory-Vocal, Mixed Mode, and Just Drive Tasks on Road:
STISIM and Sternberg Surrogates and Selected Road Driving Measures

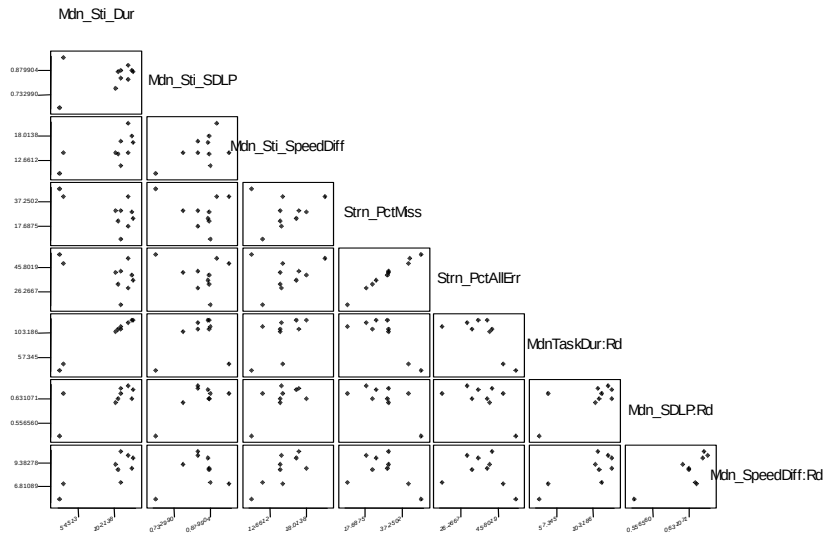


Figure 5-28. Laboratory Surrogate Correlations With Selected Road Driving Metrics

Figure 5-29 illustrates the scatter plots among several laboratory OED surrogates and road FVTS and CHMSL miss rates. The Sternberg Combined Decrement score is as good or better than other measures to predict FVTS miss rates on the road. Road CHMSL miss rates are correlated best with Sternberg Percent All Errors. However, Voice Dial (upper outlier) and Route Instruct (lower outlier) are not well represented by the trend of the remaining tasks (Figure 5-30).

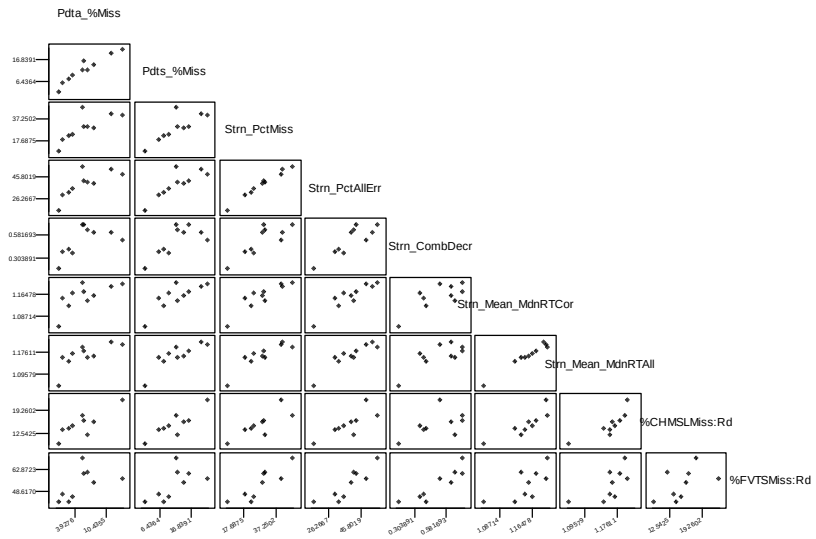


Figure 5-29. PDT and Sternberg Surrogates and Selected Road OED Measures for Auditory-Vocal, Mixed-Mode, and Just Drive Tasks

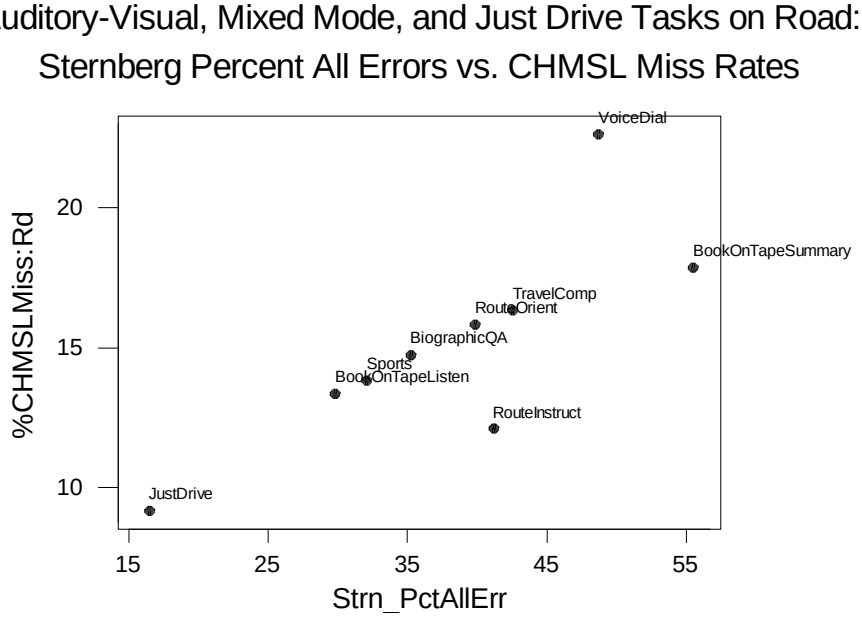


Figure 5-30. Sternberg Percent All Errors Versus CHMSL Miss Rates on the Road for Auditory-Vocal, Mixed-Mode, and Just Drive Tasks

5.4.4.1 Summary of Laboratory Surrogate Prediction with Road Auditory-Vocal, Mixed-Mode, and Just Drive Tasks

The auditory-vocal, mixed-mode, and Just Drive tasks had the following prediction profile. OWL and Multitasking Difficulty scales were not well correlated with any of the selected road measures. No surrogates were significantly associated with road driving measures. Sternberg measures were correlated best with road FVTS and CHMSL miss rates.

5.5 Summary and Recommendations for Laboratory Surrogates

This section presented details of surrogate measure repeatability, task effects, discriminability, and prediction. These results are summarized and recommendations are provided.

5.5.1 Summary

Some comments about the analysis methods used in this chapter are in order before proceeding further. Linear correlation and regression methods and scatter plots were used for split-group repeatability analysis. Other methods could have been used, e.g., paired comparison tests of equal means or medians for Group 0 and Group 1. This was not done because the emphasis of the analysis was consistency of task ordering across the two groups. This, in turn, was driven by a belief that workload assessment is an ordinal measurement process, not an absolute process. Equality of means or medians for Group 0 and Group 1 for a single task was of less interest.

Prediction was assessed with correlations and scatter plots based on small numbers of cases. It may seem imprudent to use correlation and regression methods with such small numbers of cases. This is a legitimate concern. For example, it becomes difficult to see the true pattern in the data if given one or more extreme cases that deviate from a small handful of cases to define a general trend. However, each of the cases correlated was not from a single performance or even an average of several performances by a single person. Instead, each case was from a sample of from roughly 50 persons (for laboratory or track data) to up to 100 persons (for road data), each of whom had a score that was typically based on the average of two (sometimes more) repetitions. This suggests that the stability of the data points may be greater than might be expected otherwise.

A split-group method was developed and applied to assess repeatability. Task-level summary statistics were computed separately and then correlated for two randomly selected participant subgroups approximately matched for age and gender. Most of the surrogate measures were repeatable, defined as a correlation of about 0.70 or higher between Group 0 and Group 1 results. Scatter plots and regression lines provided additional insights. Moderate to high variability about the regression line was often present. On the other hand, surrogate measures related to task time (Static Time, TSOT, and STISIM Task Duration) were exceptionally tightly clustered along a straight line. Task time values were for visual-manual tasks whose durations were inherent to the task itself. The length of a task was not arbitrarily set for these tasks. This result implies that task duration is largely limited or constrained by the nature of such tasks (number of steps, accuracy requirements, etc.) and is less subject to psychological effort or other sources of performance variation.

Mean values were sometimes correlated better across Group 0 and Group 1 than medians. This was indicated by a smaller standard error about the regression line than that found with medians. In other instances, there was no appreciable difference in the correspondence between mean and median values. Medians were subsequently used because medians are more resistant to extreme values or outliers present in the data. If there are no extreme values or outliers, the difference between mean and median should be very small.

An interesting pattern of OED miss rate results emerged. As the participant became increasingly loaded by concurrent tasks, the repeatability of the measures improved. PDT alone showed slightly less variability than PDT in STISIM. In turn, the Sternberg procedure produced even better correspondence. Increased consistency appears to come with more demands placed on the performer.

There were a few unrepeatable (or insufficiently repeatable) surrogate measures. The Situational Awareness (SA) Magnitude Estimates scale was not repeatable. The SA ratings were collected once only due to scheduling limitations. Perhaps combining reps would have resulted in more stable results. Another possibility is that the judgments required of the test participants were hard to make. A few Sternberg measures were also not sufficiently repeatable. These measures included Reaction Time (RT) for Incorrect Responses and Percent Error Given a Detection.

Task effects were displayed graphically and reported in the N-point summaries contained in Appendix Q to this report. The results indicated a paradox wherein a shorter task tended to have worse detection performance. (Similar findings were reported by Young and Angell, 2003.) Why this might arise is discussed elsewhere in this report.

Discriminability refers to the sensitivity of a measure to distinguish between two tasks, one of which is a higher-workload task and the other of which is a lower-workload task, based on prior prediction (See Chapter 2). The results showed that most measures were able to distinguish task effects from Just Drive or just performing the surrogate task alone. Beyond this, an interesting pattern of results emerged that distinguished visual-manual from auditory-vocal tasks. Visual-manual tasks were distinguished well by time-driven measures such as Static Time, TSOT, and various STISIM task duration and driving measures. Subjective workload ratings (OWL, Multitasking Difficulty) were also very good to distinguish lower- from higher-workload tasks. On the other hand, OED measures were not discriminating. This result arose from the "more is more workload" prior hypothesis that was applied to these data. The paradox that shorter duration tasks were worse on detection performance manifested itself here.

Auditory-vocal and Just Drive tasks had a different pattern of results. These tasks were not well discriminated by STISIM task duration or driving performance measures. They were well distinguished by OED measures. The difference between vehicle measures and OED results may be explained by the selective withdrawal of attention (Tijerina, 2000; Brown, 1994). Eye glances away from the road scene reflect a general withdrawal of attention from the road. That is, the visual demands of a task can influence both vehicle control and visual OED performance. Auditory-vocal tasks do not require eyeglances away from the road scene. This allows highly practiced driving skills like lanekeeping to remain largely unaffected. On the other hand, OED performance may be selectively affected by cognitive load. SpeedDiff, though able to distinguish tasks from Just Drive, nonetheless could not distinguish well between tasks of categorically different workload levels.

A surrogate's prediction was sometimes dependent on the venue as well as the task type. The following summarize the results from track and road.

- For visual-manual tasks, OWL and the Multitasking Difficulty subjective workload ratings were correlated with selected track driving and eyeglance measures. On the road, OWL and Multitasking Difficulty ratings were highly correlated with task-related glance counts and Percent Lanex Cross Trials, respectively.
- For visual manual tasks, Static Time and TSOT were both good predictors of eyeglance counts and total eyes-off-road time on the track. Both were good predictors of Task Duration on the track, as well as of SpeedDiff. TSOT was

slightly better correlated than Static Time. The R-Metric was not correlated with any track measures. On the other hand, Static Time and TSOT measures were distributed over a small range on the road because only a subset of tasks were used on the road. This created nearly vertical trend lines with relevant road measures.

- For visual-manual tasks, STISIM measures were very good predictors of comparable track driving measures. These measures included Task Duration, SDLP, and SpeedDiff. On the other hand, Percent Lanex Cross trials on the test track were not predicted by STISIM lane exceeds. STISIM Percent Lanex Cross trials was a good predictor of SDLP on the track. On the road, STISIM measures had a similar problem as the Static Time and TSOT measures. However, STISIM SDLP performed as well or better than other STISIM measure to predict road driving performance measures.
- No track OED percent missed detections were predicted by any of the laboratory surrogates for visual-manual tasks. On the road, several Sternberg measures were correlated with CHMSL miss rates for visual-manual tasks.
- The auditory-vocal, mixed-mode, and Just Drive tasks on the track were different than the visual-manual tasks in terms of surrogate prediction. Traditional task-related eyeglance measures were not applied due to the nature of these tasks.
- OWL and MD subjective workload scales were not adequately correlated with any track or road measures for the auditory-vocal, mixed-mode, and Just Drive tasks on the track.
- No track driving measures were predicted by any of the selected surrogates on either the road or track or the auditory-vocal, mixed-mode, or Just Drive tasks.
- On the track, the Sternberg method produced useful predictions of track FVTS miss rates. In particular, the Combined Decrement Score provided very good prediction. Sternberg measures were correlated best with road FVTS and CHMSL miss rates.

5.5.2 Recommendations

Table 5-11 presents the laboratory surrogates recommended for inclusion in a driver workload assessment toolkit. This table combines the results of the repeatability, discriminability, and predictive validity analyses presented earlier in this chapter into a single summary table. The complete toolkit is presented in Chapter 8. The toolkit material in Chapter 8 includes not only a description of each surrogate but also recommendations for the selection of tools for the auditory-vocal and visual-manual task types and suggestions on the selection of tools for use at various stages in the product development cycle.

As can be seen in the left-hand side of Table 5-11, all surrogates are repeatable. In addition, the surrogates generally discriminate levels of workload, although there are differences between discriminability for visual-manual and auditory-vocal tasks for some surrogates. The discriminability summary is shown in the third and fourth columns of the table. The surrogates' prediction to driving performance metrics is shown on the right side of the table. This material distills the correlation results presented in Table 5-7, Table 5-8, Table 5-9, and Table 5-10 into a qualitative presentation of a surrogate's predictive validity. In the table, a single asterisk is used to represent "good" prediction between a surrogate and a driving performance metric. Two asterisks indicate "better" correlation. The terms "good" and "better" are used here in a relative

sense. The prediction information in Table 5-11 also summarizes across the auditory-vocal and visual-manual task types and consolidates the results from test track and on-road venues.

It should be noted that some entries in the right side of the table contain empty cells. This indicates that a surrogate did not predict driving performance particularly well. An example of this is the OWL metric, which although repeatable and discriminable, did not predict driving performance. The value in including OWL resides in its reflection of driver perceptions. Other metrics with low correlations to driving performance were included in the set of recommended surrogates but for other reason. These issues are discussed in the recommended toolkit section of Chapter 8.

Table 5-11. Recommended Laboratory Surrogates

	Repeat-ability	Discriminability		Prediction for Driving Performance, Object-and-Event Detection, and Selected Eyeglance Behaviors						
DWM Surrogate		Visual-Manual Tasks	Auditory-Vocal Tasks	Task Time While Driving	Driving: SDLP	Driving: Percent Lanex (Cross)	Driving: Speed Difference	Object & Event Detection	Task-Related Number of Glances	Task-Related Total Duration of All Glances
Mean OWL	Yes	Yes (1)	No							
Multitasking Difficulty Scale	Yes	Yes (1)	No			**				
Static Time	Yes	Yes (1)	N/A (3)	*					*	*
TSOT	Yes	Yes (1)	N/A (3)	**			*		**	**
Median STISIM Duration	Yes	Yes (1)	No	**			**		*	*
Median STISIM SDLP	Yes	Yes (2)	No		**	*				
STISIM Percent Lanex Cross Trials	Yes	Yes (1)	No							
Median STISIM SpeedDiff	Yes	Yes (1)	Yes (2)				**			
PDT Alone (PDTA) Pct Miss	Yes	Yes (2)	Yes (1)							
PDT-STISIM (PDTS) Percent Miss	Yes	Yes (2)	Yes (1)					*		
Sternberg Percent Missed Detects	Yes	Yes (2)	Yes (2)					**		
Sternberg Percent All Errors	Yes	Yes (2)	Yes (1)					**		
Sternberg Median All RT	Yes	Yes (2)	Yes (2)					*		
Sternberg Combined Decrement	Yes	Yes (2)	Yes (1)					**		

Notes:

1 = Discriminates higher-workload tasks from lower-workload tasks
2 = Discriminates multitasking from Just Drive, but not higher-workload tasks from lower-workload tasks.
3 = Occlusion and Static Time surrogates were not run with auditory-vocal tasks.
* is good; ** is better

5.6 Chapter References

Asoh, A., Uno, H., Noguchi, M., and Kawasaki, Y. (2002). *Study on the static test method to evaluate the total glance time for car navigation systems while driving*. Japanese Automobile Research Institute (JARI). (Abstract).

Alliance of Automotive Manufacturers (AAM). (2003). *Statement of principles, criteria, and verification procedures on driver interactions with advanced in-vehicle information and communications systems* (Version 2.1). Southfield, MI: Author

Brown, I. D. (1994). Driver fatigue. *Human Factors*, *36*(2), 298-314.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical methods for data analysis*. Belmont, CA: Wadsworth.

Conover, W. J. (1999*). Practical nonparametric statistics* (Third edition). New York: John Wiley and Sons.

Edwards, A L. (1984). *An introduction to linear correlation and regression*. San Francisco: W. H. Freeman and Company.

Fleishman, E., and Quaintance, M. (1984). *Taxonomies of human performance: The description of human tasks.* New York: Academic Press.

Gescheider, G. A. (1997). *Psychophysics: The fundamentals* (Third edition) (pp. 231-263). Mahwah, NJ: Lawrence Erlbaum.

Green, P. and Tsimhoni, O. (2001). *Visual occlusion to assess the demands of driving and tasks: the literature*. Presentation at the Exploring the Occlusion Technique: Progress in Recent Research and Applications Workshop, Torino, Italy. Available at: www.umich.edu/~driving/occlusionworkshop2001

Hart, S. G., and Wickens, C. D. (1990). Workload assessment and prediction. In H. R. Booher (Ed.), *MANPRINT: An approach to system integration* (pp. 257-296). New York: Van Nostrand Reinhold.

Hartwig, F., and Dearing, B. E. (1979). *Exploratory data analysis*. Thousand Oaks, CA: Sage Publications.

Hill, S. G., Iavecchia, H.P., Byers, J.C., Bittner, A. C., Jr., Zaklad, A., and Christ, R. (1992). Comparison of four subjective workload rating scales. *Human Factors*, *34(4)*, 429-440.

ISO (International Standards Organization). (2004). *Road vehicles – Ergonomic Aspects of transport information and control systems—Occlusion method to assess visual distraction due to the use of in-vehicle systems* (ISO/CD 16673).

Kroemer, K. (1986). *Engineering physiology*. Amsterdam: Elsevier.

Lane, N., Kennedy, R. S., and Jones, M. B. (1986). Overcoming unreliability in operational measures: The use of surrogate measure systems. *Proceedings of the Human Factors Society 30th Annual Meeting*, 1398-1402.

Lodge, M. (1981). *Magnitude scaling: Quantitative measurement of opinions*. Beverly Hills, CA: SAGE Publications.

Marascuilo, L. A., and McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks/Cole Publishing Co.

SAE (Society of Automotive Engineers). (2004). *J2364 navigation and route guidance function accessibility while driving*. Warrendale, PA: Society of Automotive Engineers.

Smith, P. J., and Langolf, L. (1981). The use of Sternberg's memory scanning paradigm in assessing effects of chemical exposure. *Human Factors, 23*(6), 701-708.

Sternberg, S. (1998). Discovering mental processing stages: The method of additive factors. In D. Scarborough and S. Sternberg (Eds). *An invitation to cognitive science, Volume 4: Methods, models, and conceptual issues* (2nd Edition) (pp. 703-864). Boston: MIT Press.

Tijerina, L. (2000). *Issues in the evaluation of driver distraction associated with in-vehicle information and telecommunications systems*. Paper presented at the Transportation Research Board, 79th Annual Meeting, Washington, DC, January 9 - 13, 2000.

Wickens, C. D., Hyman, F., Dellinger, J., Taylor, H., and Meador, M. (1986). The Sternberg memory search task as an index of pilot workload. *Ergonomics*, *29*(11), 1371-1383.

Wierwille, W. W. (1993). Visual and manual demands of in-car controls and displays. In B. Peacock and W. Karwowski (Eds.), Automotive ergonomics (pp. 299-320). London: Taylor and Francis.

Young, R. A., and Angell, L. S. (2003) The dimensions of driver performance during secondary manual tasks. Proceedings of "Driving Assessment 2003: The Second International Symposium on Human Factors in Driver Assessment, Training and Vehicle Design," Park City, Utah. July, 2003.

6 Analytical Modeling Results

6.1 Background

Ideally, ergonomics input is factored in early in the product development process. Early introduction offers the greatest flexibility to consider design alternatives. Considering ergonomic input later in product development provides fewer opportunities for design changes.

Analytical human performance models also can impact product design early in the product development process. These models can be used to estimate the performance impacts of different tasks or design alternatives. In addition, they can be used before interactive prototypes are available for human subjects testing. Well-established models (e.g., legibility models) can also be used instead of human subjects testing.

Based on the results of an extensive task analysis, the DWM project team developed an analytic modeling methodology, the Task Steps Assessment, with the following four outputs:

- **Count of Task Steps** sums the number of task steps or subtasks required to complete each task. The simple notion of this output was that a task or design with fewer steps is generally better. The predictor score is the total number of task steps, however individual counts of physical and cognitive steps also may offer insight into the task, especially for certain classes of tasks.
- **Activity Time** estimates the time required for each task step and sums these times. The idea behind this model is that not all task steps are equivalent and shorter tasks times are generally better. The predictor score is the total sum of times assigned to all physical and cognitive steps. Alternate statistics that may offer information about the task are separate physical and cognitive times and mean step times.
- **Dual Task Conflict Potential** estimates the competition for required driver sensory input, working memory, and response resources between the driving task and a concurrent secondary task and is based on Modified Multiple Resource Theory (Modified MRT). The idea behind Modified MRT is that less resource competition between concurrent tasks is generally better. The predictor score is the Dual Task Conflict Potential (DTCP) as calculated from a resource vector similar to Modified MRT. See Appendix A and Task 1 Literature Review for more information on Modified MRT and DTCP.
- **Activity Time Weighted Dual Task Conflict Potential** combines the Activity Time and the DTCP in an attempt to provide a more comprehensive estimate of each task step's demand on the driver. The predictor score is the total sum of physical and cognitive activity times for each task step weighted by the DTCP value for that step.

None of the model outputs described in this chapter used CAMP road, track, or laboratory surrogate data so as to avoid circularity. The modeling methodology was developed from data and models in the published literature. The models are, in this sense, truly analytical because they do not need data from human subjects testing of the system or task being evaluated.

Another modeling tool that could have been explored in the CAMP DWM project was the In-Vehicle Information System Design Evaluation and Model of Attention Demand (IVIS DEMAnD) from (Hankey et al, 2000a). The IVIS DEMAnD was developed at Virginia Polytechnic Institute and State University (Virginia Tech) under contract to the Federal Highway Administration (FHWA). Currently, the model exists only in a pre-prototype state due to

termination of funding after the initial stage of development. It is intended to aid designers of IVIS products to predict driver workload levels as a function of design characteristics. It makes use of empirical data from numerous studies conducted at Virginia Tech and elsewhere. Rather than containing a model of driving, per se, it appears to contain tables of driving performance regression models that are indexed by various human-machine interface tasks.

The IVIS DEMAnD model was not selected for use in this project for several reasons. First, its successful application depends heavily on the empirical database and library of tasks upon which it was developed. Both of these (as they exist in the original pre-prototype state delivered to FHWA) were not sufficient to model the full set of tasks used in the DWM project. While the model is extendable and offers two modeling mechanisms that permit additional data to be incorporated, the constraints of the DWM project precluded the development or adaptation of an extended version of the model. Therefore, it was not tractable to apply the model to the particular tasks used in the DWM project as an analytical tool. In addition, the CAMP DWM project team was particularly interested in analytic tools that did not depend upon and were not derived from driving performance data obtained from this project, but were based instead upon only attributes of the task and its structure and/or upon empirical data about human performance independent of driving per se. Since IVIS DEMAnD is a model based on empirical driving performance data, it did not meet this last criterion set by the DWM project team as desirable.

Although the IVIS DEMAnD model was not selected for use in this project, there is at least one automotive company that has found IVIS DEMAnD, when coupled with an extended and continuously evolving empirical database and library of task elements, to be a useful analytical tool for predicting driving performance metrics early in development of in-vehicle systems. This combination yields moderately high to high correlations with test data that is subsequently collected under test track conditions for the same in-vehicle systems. So the fact that IVIS DEMAnD could not be included in the DWM work does not necessarily mean that it should be ruled out as a useful approach to analysis in this domain. It was simply a tool that did not meet the needs of this particular application.

6.2 Method

6.2.1 Analysts

Four members of the DWM project team modeled the DWM tasks using the Task Steps Assessment method. All four analysts were trained in human factors, but had differing levels of education, widely varying amounts of experience (between 5 and 30 years), and varied backgrounds.

6.2.2 Procedure

Task analysis and modeling was completed in two different stages, a development stage followed by an implementation stage. Two analyst-modelers were involved in the development stage, as indicated above. These analysts are engineers with human factors backgrounds including differing amounts of experience with modeling human performance. The analyst-modelers were also experimenters who conducted the actual laboratory and in-vehicle testing of subjects performing the CAMP study tasks. This combination provided well-trained individuals possessing a very detailed knowledge of not only the tasks but also how different people actually performed the tasks.

The modelers started with an initial task analysis instruction manual and preliminary verb sets prepared for this project. An initial set of elemental activity times was also prepared for physical and cognitive activity time predictions associated with the verb set. As tasks were modeled and weaknesses in the system were identified, the verb lists were augmented to make as complete a system as possible for modeling human task performance. This does not mean that the initial modelers collaborated on the actual construction of their models. Rather, the collaboration was an effort to improve verb definitions and explanations and the addition of new verbs to the sets to allow for the tasks being modeled. The improved Verb Lists and Elemental Activity Times are presented in Table 6-1 through Table 6-4.

The initial analyses lead to the development of three spreadsheet tools, a revised instruction manual, and example task models. The spreadsheet tools are based on Visual Basic macros running in Excel spreadsheets that provide a pull-down, menu-based graphical user interface to construct the models with a uniform structure.

- The first of the spreadsheet tools allows a modeler to enter the task steps, descriptions, concurrency, and a vector for the Modified MRT resources required for each task step.
- The second tool allows the modeler to pass through the models as constructed and assign time values to each step. These time values are based on work and research by the DWM team and provide an estimate of the time required to complete a task. This tool also calculates a number of metrics based on step counts and times.
- The third tool uses the output of the other two to compute a number of metrics based on Modified MRT and generates a summary set of metrics for each model.

An instruction manual detailing the methodology, verb lists, and activity times for the modeling method was prepared next. This manual included three example tasks, modeled by the two initial modelers. These tasks were invented tasks that one might perform in a car and did not bear direct resemblance to the DWM tasks to be modeled. Instructions were also included on how to use the spreadsheet tools to develop the models. To ensure modelers were examining the same tasks, as there were numerous versions of each task, a template was developed and used. The template specified the exact task settings, and where applicable, the text of task recordings in order to provide a uniform structure for each modeler's assumptions and comments.

After completion of the methodology and tools development and the initial modeling efforts, two new modelers, who were previously unexposed to any of the task step assessment development, were recruited. Like the first two modelers, the two new modelers varied greatly in education, experience, and training. They were given the same materials and introduction to the methodology and then were asked to independently model the tasks. The instruction manual, which experimenters used to conduct the tasks with subjects previously in the project, provided details about the tasks and the experimental method that had been used. Thus both modelers had the same set of instructions for the tasks and the modeling methodology.

Table 6-1. Verb List for Physical Actions

(Based Loosely on MTM-1 Conventions)

Reach: • Definition: To transport the hand or fingers to a destination. • Distinction: **Reach** is distinct from **Move** in that the primary objective is transport of the hand or fingers from an initial state to a device or an object rather changing the location of the object. • Example: To **Reach** for a cell phone in the center console cup holder.
Move: • Definition: To transport an object to a destination. • Distinction: **Move** is distinct from **Reach** in that the primary objective is transport of an object rather than to transport the hand. • Example: **Move** a cell phone to the ear.
Turn: • Definition: To rotate the empty or loaded hand about the long axis of the forearm. • Example: **Turn** a key already inserted into the vehicle ignition.
Press-Momentary (apply pressure): • Definition: To apply muscular force to overcome object resistance, accompanied by little or no motion. • Distinction: This is a momentary push button activation and a **Release** step is implicit rather than a separate later step. • Example: **Press** a key on a keyboard (e.g. typing).
Press-Hold (apply pressure for some duration): • Definition: To apply muscular force to overcome object resistance, accompanied by little or no motion for some period of time. • Distinction: This is a prolonged application of pressure and *does* require an explicit **Release** step at some later point in time. • Example: To hold down a scroll button to move through a list on a display screen.
Grasp: • Definition: To lay hold of or clasp by means of the fingers or arms. • Distinction: **Grasp** is distinct from **Move** in that an object must first be grasped before it can be picked up. • Example: **Grasp** the volume control knob on a car radio.
Position: • Definition: To bring one object into an exact predetermined relationship with another object. • Example: **Position** (locate) a key into the vehicle ignition.
Release: • Definition: To un-grasp or relinquish hold of an object, usually by opening the fingers. • Example: **Release** the cell phone onto the center console cup holder.
Look: • Definition: Move the gaze (i.e., the cone of fine resolution vision) from one location to another known location (target or destination). • Clarification: **Look** is distinct from **Search** in that the destination location is known when looking but is not known initially in **Search**. Look is also used for a unique or prominent location or feature such as a large isolated radio button. • Example: **Look** to the Message Center Display or a radio-tuning knob.

Table 6-2. Verb List for Cognitive Activities

(Based on Literature Review)

Read-Text: • Definition: To interpret the meaning of visually displayed alphanumeric material such as numbers, words, and text. • Example: **Read** an email or printed text.
Read-Iconographic: • Definition: To interpret the meaning of visually displayed iconographic information such as signs, symbols and indicator lights. • Example: Examining an Instrument Panel Telltale Indicator for oil pressure.
Listen: • Definition: To interpret the meaning of auditorily presented verbal material such as numbers, words, and text. • Example: **Listen** for the desired function in a telephony display.
Speak: • Definition: To utter words or articulate sounds with the voice. • Example: To say aloud a response to a biographical question.
Write: • Definition: To form alphanumeric characters with a pen, pencil, or other hand-held instrument. • Example: **Write** a letter on the PDA screen.
Check: • Definition**:** Visually evaluate the state of a display. • Example: **Check** a speedometer to see if you are traveling 'around' the speed limit.
Compare: • Definition: Examine similarities and differences between two or more objects or events including the translation of a concept into the object or event to be compared to. • Distinction: This operation includes any mental processing needed to translate feelings, concepts, icons etc. into homogeneous forms that can be directly compared. • Example: Comparing cold feet to the position of HVAC control knob icons.
Search-visual: • Definition: To visually scan for one or more target items among a set of non-target items. • Distinction: A search is a look to an unknown target, which requires scanning of an area to locate a target, which is not distinct or prominent. • Example: Locating a Point of Interest in a list or a button in a group of buttons.
Select: • Definition: **Select** from among two or more alternatives. • Example: **Select** the third preset button on the car radio.
Track: • Definition: Continuously monitor a moving target and adjust a control to maintain target-vs.-signal error within acceptable limits. • Example: **Track** the lane width while driving.
Calculate mentally: • Definition: To carry out arithmetic processes (addition, subtraction, multiplication, division) without pencil and paper, i.e., 'in the head'. • Example: To add the toll amounts that will be required for a trip.
Spatial Calculate: • Definition: To carry out spatial based calculation without pencil paper. • Example: Determining the relative size of groups of objects.

Rotate mentally: • Definition: To mentally revolve or turn about an axis or direction. • Example: To rotate an upside down letter in a word of text to read the word.
Adjust: • Definition: To mentally plan a transformation of an object from a start state to an end state. • Example: Determining movement required to set a control to a desired position.
Recall-STM: • Definition: Retrieve information from working memory. • Example: To **Recall** the outcome of a game after hearing a sports broadcast.
Recall-LTM: • Definition: Retrieve information from long-term memory. • Distinction: This **Recall** is used when a search must be made of long-term memory to produce the required information. • Example: To **Recall** the name of a co-worker at a previous job.
Feedback-Auditory: • Definition: To perceive a non-verbal auditory feedback from a control. • Example: To hear a cell phone beep resulting from a button press.
Feedback-Tactile: • Definition: To perceive a tactile feedback from a control. • Example: To feel the detent click between positions of an HVAC control knob.
Retain: • Definition: Retain information in working memory. • Example: To remember the series of directions to get to a destination.
Wait: • Definition: To **Wait** for a system response before beginning the next step in a sequence that is externally paced. • Example: To **Wait** for a phone to display test indicating it is ready for a call.
Determine: • Definition: A GOMS-like “bypass” step that takes the place of a complex mental process that may include multiple cognitive steps like **Read**, **Calculate** and **Adjust** but that cannot be fully modeled with this vocabulary, not a commonly used task step. • Example: Determining the missing word in a passage of text, which includes **Compare**, **Recall** and **Read** but is not fully explained by these steps alone.

Table 6-3. Physical Activity Time Models

(Modeled after MTM-1)

Activity	Case	Activity Time Estimates Let y = Activity Time Estimate, in seconds x = Distance, in inches, or x = Degrees of Rotation (as needed)
Reach	**A**: Reach to object in fixed location or in other hand or on which other hand rests	$y = -0.0002x^2 + 0.0236x + 0.0988$ $R^2 = 0.9878$
	B: Reach to a single object in a location that may vary slightly from cycle to cycle	$y = -0.0003x^2 + 0.0355x + 0.0806$ $R^2 = 0.9959$
	C: Reach to object jumbled with other objects in a group so that search and select occur	$y = -0.0003x^2 + 0.0364x + 0.1279$ $R^2 = 0.9884$
	D: Reach to a very small object or where accurate grasp is required	$y = -0.0003x^2 + 0.0364x + 0.1279$ $R^2 = 0.9884$
	E: Reach to an indefinite location to get hand in position for body balance, or next motion, or out of the way	$y = -0.0002x^2 + 0.0314x + 0.0873$ $R^2 = 0.9931$
Move	**A**: Move object to other hand or against stop	$y = -0.0002x^2 + 0.0352x + 0.0699$ $R^2 = 0.9982$
	B: Move object to approximate or indefinite location	$y = -0.0004x^2 + 0.0372x + 0.0914$ $R^2 = 0.9935$
	C: Move object to exact location	$y = -0.0003x^2 + 0.0406x + 0.1054$ $R^2 = 0.9949$
Turn	**Turn**: Assume negligible resistance and x degrees of rotation	$y = 0.0016x + 0.0537$ $R^2 = 0.9998$
Press-Momentary	**Momentary** (e.g., keying time for a simple pushbutton) from (Munger, Smith, and Payne, 1962)	y = 0.63
	Keystroke random words (not prose), single key stroke time from (Card, Moran, and Newell, 1983)	y = 0.25
Press-Hold	**Hold** (determined by design requirement)	y = [design HOLD duration}
Release	**Release**: Normal release by opening fingers	y = 0.07
Position	**Loose:** (approximate, no pressure needed) (Distance to engage is $\leq$ 1 inch for all cases)	y = 0.20
	Close: (light pressure needed)	y = 0.58
	Exact: (tight fit, heavy pressure needed)	y = 1.55
Grasp	**A**: Any size object by itself, easily grasped	y = 0.07
	B: Object very small or lying close against a flat surface	y =0 .13
	C1: Diameter > 0.5 inch; interference with grasp on bottom and one side of nearly cylindrical object	y = 0.26
	C2: ¼ inch < Diameter < ½ inch; same interference as C1	y = 0.31
	C3: Diameter < ¼ inch; same interference as C1	y = 0.39
	Re-grasp or change hands without loss of control	y = 0.20
	4A: Select 1"x1"x1" or larger object jumbled with others	y = 0.26
	4B: Select ¼"x ¼ "x 1/8 " to 1"x1"x1" object, jumbled	y = 0.33
	4C: Select smaller than: ¼"x ¼ "x 1/8 ", jumbled	y = 0.46
Look	**Look**: T=distance in inches from look location to the next; D= viewing distance in inches perpendicular to the line of travel	y = 0.55*(T / D)

Table 6-4. Cognitive Activity Time Models

(Literature Sources in Parentheses)

Activity	Input Description	Activity Time Estimates Let y = Activity Time Estimate, in seconds
Read-Text	**Text**: For connected prose, assume 214 words/minute reading rate or 9th grade reading level (from Crowder, 1982).	y = 0.28x For x = number of words
	Label: For a single word in isolation, assume single word or label read in a single saccade of 0.288 seconds duration (from Card, 1983).	y = 0.288
Read-Icon	**Icon**: Assume an icon is read in a single saccade of 0.288 s.	y = 0.288
Listen	**Maximum** transmission rate for speech is 250 words/min (from Deatherage, 1972).	y = 0.24x for x = number of words heard
	Relaxed transmission rate for speech is 150 words/min (from Ericsson and Simon, 1984).	y = 0.40x for x = number of words heard
Speak	**Maximum** transmission rate for speech is 250 words/min (from Deatherage, 1972).	y = 0.24x for x = number of words heard
	Relaxed transmission rate for speech is 150 words/min (from Ericsson and Simon, 1984).	y = 0.40x for x = number of words heard
Write	**Write**: The maximum writing rate of 100 characters/ minutes is equal to 0.6 seconds per character (from Devoe (1967) as cited in Seibel, 1972).	y = 0.6x for x = number of characters written
Check	**Read-Label**: Use Read-Label activity time for a check glance.	y = 0.288
	Feedback-Visual: Estimate with simple reaction time of 200 mseconds (from Frost, 1972, Fig 6-60).	y = 0.200
Compare	**Easy**: Trabasso (cited in Carroll and Freedle, 1972) reported a time of .45 seconds. for verifying stored information mentally and .08 seconds. to compare that information to other objects	y = 0.53
	Hard: Trabasso (cited in Carroll and Freedle, 1972) reported a longer time of 1.24 seconds. for verifying stored information mentally and 0.27 seconds. to compare that information to other objects.	y = 1.51
Track	**User Enter**: Either track until target is at rest or until target is lost.	User determined
Calculate mentally	**Default**: Based on the common occurrence of mental operation times in the area of 1.2 seconds and the reported J2365 mental time this step is assigned a fixed value of 1.2 seconds.	y = 1.2
Calculate spatially	**Default**: Based on the common occurrence of mental operation times in the area of 1.2 seconds and the reported J2365 mental time this step is assigned a fixed value of 1.2 seconds.	y = 1.2
Rotate mentally	**Rotate Mentally**: Shepard and Metzler (cited in Wickens and Hollands, 1999) reported mental rotation rates for rotations in the picture plane of approximately 60 degrees/sec.	Y = 0.0167x for x = degrees of mental rotation
Recall-STM	Default: Card, Moran, and Newell (1983) present mental times ranging from 0.62 seconds to 1.35 seconds, with the latter including a choose operation. Recall-STM is assigned a time of 0.62 seconds here based on the quicker operation above.	y = 0.62
Recall-LTM	Default: A value of 1.35 is assigned for Recall-LTM as it is logical that it takes a longer period and may involve a selection or choose operation as the mental operation in Card, Moran and Newell (1983).	y = 1.35
Adjust	Reported standard mental times range from 0.62 seconds to 1.2 seconds. Familiar: this simple mental adjustment uses the quicker reported mental time of 0.62 seconds.	y = 0.62
	Complex: this more difficult version of mental adjustment uses the longer mental time of 1.2 seconds.	y = 1.2
Feedback-Auditory	Feedback-Auditory: Estimate with simple reaction time of 160 mseconds to moderate-intensity auditory stimulus from Frost (1972).	y = 0.16

Activity	Input Description	Activity Time Estimates Let y = Activity Time Estimate, in seconds
Feedback-Tactile	Feedback-Tactile: Estimate with simple reaction time of 150 mseconds to moderate-intensity tactile stimulus from Frost (1972).	y = 0.15
Retain (In WM)	Default A study by Trabaso (cited in Carroll and Freedle, 1972) offers a time range of 0.56 to 1.28 seconds for storing information in memory (and includes some mental transformation). Assuming that some mental operation is required in conjunction with the retention phase this activity is assigned a fixed time of 1.2 seconds.	y = 1.2
Select (Decide on an option)	Decide: Use Hick's law with n+1 alternatives. The logic of adding 1 is it deals with the additional alternative of whether to select or not. Multiplier of 0.15 is from Card, Moran and Newell (1983). Bits (log to base 2) are used in calculation and calculated as log(x)/log(2).	y = 0.15 * [log(x+1)/log(2)] for x = number of alternatives
Wait	Wait: Defined by system response time or else unconstrained.	Mean system response time
Determine	Default: Step is a combination of at least 4 mental operations, thus uses 4*(J2365 'mental' time of 1.2 seconds)+ (reaction time of .2 seconds) = 5 seconds.	y = 5.0
	User Enter: Enter estimate of time for the specific situation.	User specified value
	Problem Solving Aloud: Ericsson and Simon (1984) reported average rate of 40 to 72 words per minute for spoken problem solving. An average of this range yields 1.165 seconds per word for problem solving out loud activities.	y = 1.165x for x = number of words spoken aloud
	Think Aloud: Ericsson and Simon (1984) reported average rate of 50 to 110 words per minute for thinking aloud. An average of this range yields a rate of .875 seconds per word for thinking out loud activities.	y = 0.875x for x = number of words spoken aloud
	Complete Fragment: Kintsch (1974) reported values of 4 seconds. for completing either a simple or complex sentence fragment.	y = 4.0

6.3 Results and Discussion

Repeatability and predictive validity are generally addressed separately. However, the two are closely intertwined in the modeling efforts. Therefore, the results presented address both inter-analyst repeatability and predictive validity in a way intended to support a better appreciation of the findings.

The purpose of this section is to illuminate some general results of modeling as applied to the DWM task set. These are intended as observations on the methodology itself, both to aid in future application of the system as well as to offer a guide to potential improvements to the system. Detailed discussion of the results of the modeling effort as related to the specific tasks and correlation to real world testing will be presented in later sections.

6.3.1 Modified Multiple Resource Theory Modeling

The development of a modeling system with correlation to real-world performance of in-vehicle tasks is very appealing. The use of such models can reduce, but never eliminate, the need for actual testing with subjects in a laboratory or in real-world driving. Thus, a modeling methodology should be generally applicable to various task types and offer flexibility to accommodate differences between similar types of tasks or task components. The models should also contain enough detail to be of aid in the design phase of product development, prior to the construction of a prototype, after which the costs of re-design grow significantly.

A basic modeling strategy could be to implement Modified Multiple Resource Theory on a task level, i.e., to assign an entire secondary task one Demand Vector based on Input Modality, Working Memory, Output Modality, and Cognitive Processing. The DWM task set was modeled in this way according to the Modified MRT, which has its basis in Wickens (1999), and other related work by Wickens, but was developed specifically for this project by the project team (as is described in Appendix A). With this methodology a task, as a whole, is given a single demand vector. This vector is then compared with an overall demand matrix and a Total Interference Potential score is calculated. The results of this modeling are presented in Table 6-5.

Table 6-5. Predicted Total Interference Potential Values by Task

Tasks	Predicted Modified MRT TIP Value	Demand Vector*	Difficulty Levels Modeled: Easy Level	Mod. Level	Hard Level
Vocal-Manual Tasks					
10 Just Drive	------	------	------	------	------
3 HVAC**	1.524	V- V – M	1.524	1.774	2.024
24 Read Easy	1.761	V – V – V	1.461	1.711	1.961
4 Radio Tuning Easy	1.774	V- V – M	1.524	1.774	2.024
1 Coins	1.84	V- S – M	1.59	1.84	2.09
2 Cassette	1.84	V- S – M	1.59	1.84	2.09
28 Map Easy	1.857	V – S – V	1.607	1.857	2.107
25 Read Hard	1.961	V – V –V	1.461	1.711	1.961
5 Manual Dial	2.024	V- V – M	1.524	1.774	2.024
14 Radio Tuning Hard	2.024	V– V – M	1.524	1.774	2.024
16 CD Track 7	2.024	V- V – M	1.524	1.774	2.024
17 Route Tracing	2.09	V – S – M	1.59	1.84	2.09
29 Map Hard	2.107	V – S – V	1.607	1.857	2.107
21 Nav Dest Entry	2.553	V– SV – M	1.893	2.223	2.553
Mixed-Mode Tasks					
8 Voice Dial	2.487	VA-V-M	1.827	2.157	2.487
18 Delta	2.941	VA-V-MV	2.108	2.524	2.941
Auditory-Visual Tasks					
19 BOT Summary	1.39	-----V – V	1.06	1.225	1.39
11 Bio Q & A	1.398	A- V – V	1.398	1.648	1.898
9 BOT Listen	1.793	A- V –----	1.123	1.288	1.453
13 Sports Broadcast	1.898	A– V – V	1.398	1.648	1.898
6 Travel Comp	1.961	A- S – V	1.461	1.711	1.961
7 Route Orient	1.961	A- S – V	1.461	1.711	1.961
12 Route Instruct	1.961	A– S – V	1.461	1.711	1.961

* Demand Vector is coded in terms of Input Mode (V=Visual, A=Auditory, VA=Visual and Auditory), Working Memory (S=Spatial, V= Verbal, SV=Spatial and Verbal), and Output Mode (M=Manual, V= Vocal, MV=Manual and Vocal).

** HVAC could have been coded as V-S-M, but there was mixed opinion among the technical team, so it was left it as visual –verbal-manual.

This is a simplistic and easily implemented modeling strategy, however, due to long or complex tasks, this strategy does not offer the detail or flexibility required to make distinctions between similar tasks. When this type of model was correlated to the DWM surrogate, on-road, and test track measures, correlation results were mixed. Laboratory correlations are shown for both visual-manual and auditory-vocal tasks in Table 6-6. Correlations were computed at the task level.

Table 6-6. MMRT TIP Score Correlations to Lab Metrics

Predicted Modified MRT TIP Value Correlations			
	Lab		
Metric	*All*	*V-M*	*A-V*
Mean Static Task Time	0.827	0.819	
Median Static Task Time	0.831	0.824	
Mean Total Shutter Open Time	0.767	0.835	
Median Total Shutter Open Time	0.764	0.841	
Median STISIM Standard Deviation of Lane Position	0.379	0.727	0.429
STISIM Percent Cross Trials	0.512	0.782	0.446
Mean STISIM Speed Difference	0.609	0.928	0.635
Median STISIM Speed Difference	0.603	0.914	0.480
Median STISIM Task Duration	0.223	0.858	0.494

The correlations seen for laboratory metrics when the entire task set is examined only indicate that this modeling method has predictive power for the time-related metrics, Static Time and TSOT. These are two tests that were only conducted on the visual-manual task set. Thus, the task set was divided into visual-manual and auditory-vocal tasks, the mixed-mode tasks were excluded, and correlations were re-computed. This change yields correlations to several metrics other than time such as Mean STISIM Speed Difference. When examining this relationship, much of the correlation derives from the outlier task, Destination Entry as can be seen in Figure 6-1. When Destination Entry is excluded, the R^2 value drops to 0.70 for the visual-manual task set. Correlation to the auditory-vocal tasks is very low and the mixed-mode tasks do not group with either of the other groups.

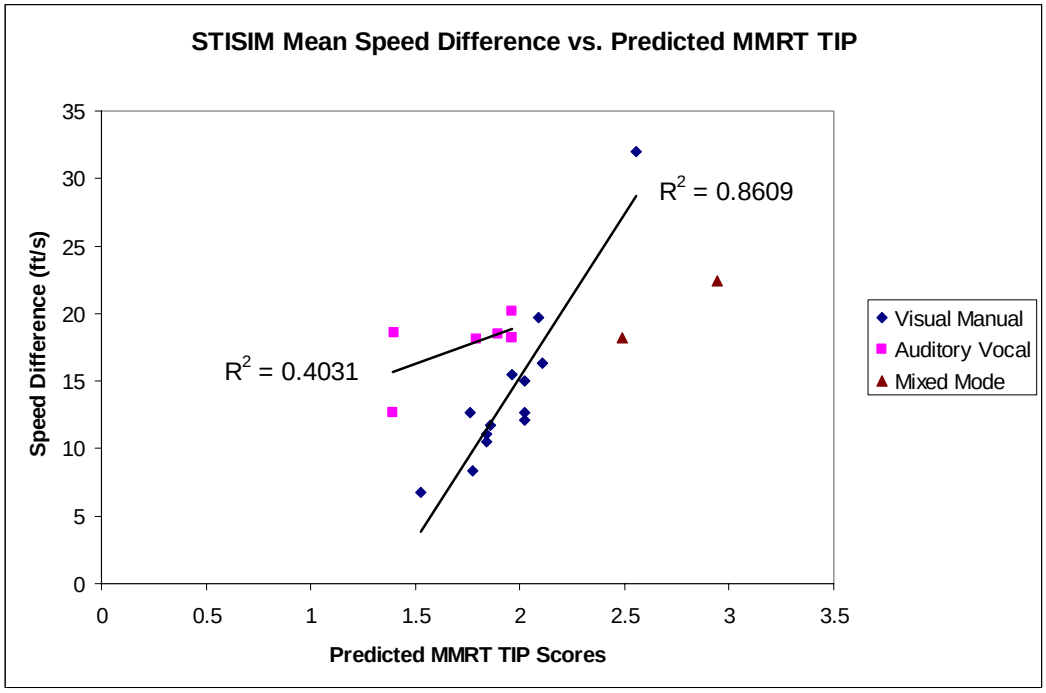


Figure 6-1 STISIM Speed Difference Versus Predicted MMRT TIP.

The Modified MRT TIP scores were also examined for correlation to on-road and test track metrics and these results can be seen in Table 6-7. Note that the definitions of the eyeglance metrics were presented in Chapter 3, Table 3-9. These results show that correlations to the task set as a whole are very low for both the on-road and test track data. When the correlations are computed by task type as was done for laboratory data correlation increased. For the on-road data only, correlation to a few eyeglance metrics become significant and three of the four are for visual-manual tasks. For the test track however, many correlations become significant for visual-manual tasks but not for the auditory-vocal tasks. Again the highest correlation is with Mean Speed Difference having an R^2 of 0.88 for the visual-manual task set. This relationship can be seen in Figure 6-2, again much of the correlation is due to the location of Destination Entry. When this task is removed from the visual-manual task set the R^2 drops to 0.50, and predictive power is reduced.

Table 6-7. MMRT TIP Score Correlations to Vehicle Metrics

Predicted Modified MRT TIP Value Correlations						
	Road	Track	Road		Track	
Metric	*All*	*All*	*V-M*	*A-V*	*V-M*	*A-V*
Mean Task Duration	-0.052	0.256	0.661	0.497	0.831	0.474
Median Task Duration	-0.057	0.225	0.624	0.501	0.830	0.495
Mean SDLP	0.279	0.522	0.559	0.471	0.816	0.547
Median SDLP	0.303	0.537	0.577	0.458	0.808	0.615
Mean Speed Difference	0.060	0.398	0.647	0.553	0.880	0.582
Median Speed Difference	0.051	0.427	0.619	0.486	0.873	0.674
Percent CHMSL MissRate	0.020	-0.170	-0.518	-0.429	-0.223	-0.791
MeanTskglncs	0.085	0.476	0.692	0.309	0.832	0.273
MeanTaskdur	-0.007	0.259	0.671	0.513	0.842	0.472
MeanglncesRD	0.072	0.456	0.685	0.308	0.833	0.279
MeanduratRD	-0.032	0.143	0.501	0.538	0.842	0.500
MeangrateRD	0.094	0.004	-0.098	-0.434	-0.053	-0.709
MeanglncesSA	-0.102	0.033	0.814	0.298	0.857	0.281
MeanduratSA	-0.098	-0.020	0.692	0.352	0.868	0.273
MeanmedSAdur	0.021	-0.201	-0.370	0.762	-0.139	0.539
MeanglncesMR	-0.108	0.061	0.875	0.263	0.837	0.264
MeanduratMR	-0.105	0.004	0.749	0.313	0.848	0.257
MeanglncesNR	0.100	0.490	0.691	0.311	0.834	0.267
MeanduratNR	0.180	0.499	0.638	0.335	0.829	0.300

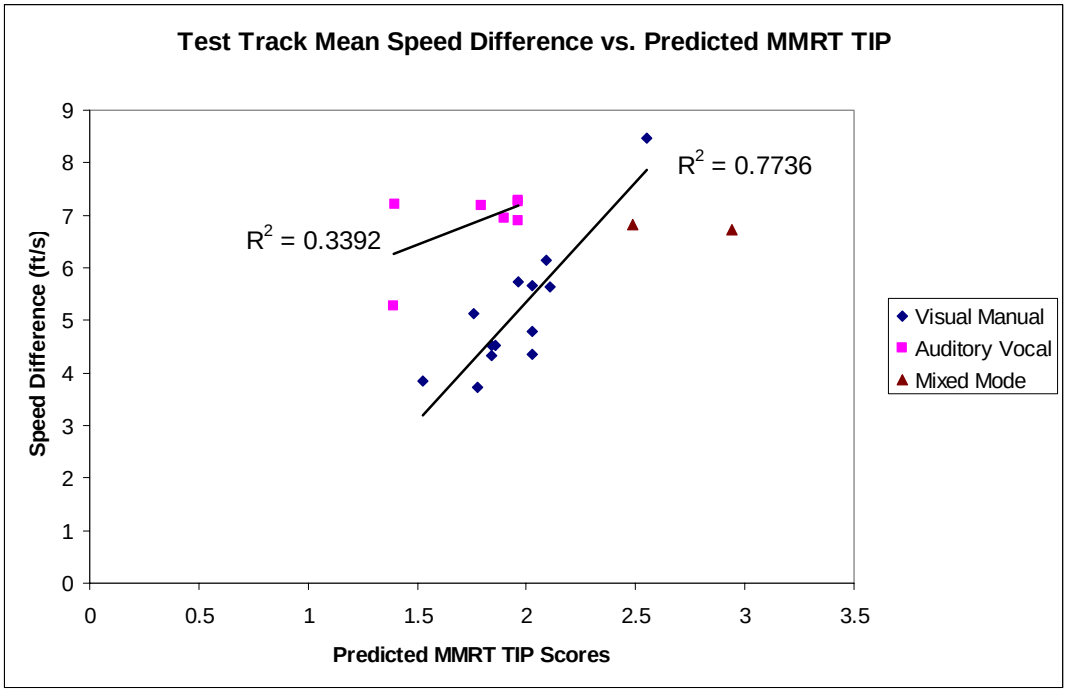


Figure 6-2. Test Track Speed Difference Versus Predicted MMRT TIP.

6.3.2 Task Steps Assessment

Given the correlations and predictive power of the Modified MRT TIP score implementation a more detailed method would seem to be required. Detailing a task by individual steps, both physical and cognitive, and then describing each step allows for more flexibility in describing and differentiating longer or more complex tasks. The output of these models is a simple count of the number of task steps. Due to numerous individual differences, however, these counts tended to be noisy and poorly correlated to real-world metrics of performance.

The addition of timing assignments to the simple step models described above yields a set of models with greater ability to distinguish tasks by yielding an activity time estimate for a task. While the times assigned to a task step are from a literature review of physical and cognitive research, they do not represent actual task completion time estimates as much as they are a measure of task demand. This is due to the complex nature of human dual-task performance and the simplicity of the modeling methodology.

To further augment the methodology and in an effort to further discriminate tasks, Modified MRT was incorporated into the models. In this way, tasks can be differentiated based on the duration of each type of demand conflict within a task. The metric yielded by this addition is the sum of the potential conflict for all steps in a model. This completed model, with task step counts, times, and conflict potentials is the result of the Task Steps Assessment Methodology.

With the complete sets of models from three modelers described in the previous section, various analysis were conducted to see if the end product was detailed and flexible enough to distinguish tasks in correlation to real world performance data. The results of a simple correlation analysis between select mean metrics yielded by the set of models and on-road data can be seen in Table 6-8. In this table, correlations greater than 0.707 (i.e., relating to an R^2 greater than 0.50) are highlighted yellow and the column labels corresponding to the numbered metrics are listed at the top of the left-most column.

Scanning down the columns labeled 1 and 2 (1 = Step Count and 2 = Step Count Rank) in Table 6-8, it becomes evident that the simple step count generated by the first part of this modeling methodology offers little correlation to real world metrics. This contrasts with the findings of Angell, Young, Hankey, and Dingus (2002), who reported high correlations between number of steps in a task and the driving performance measures of lane deviations and speed deviations. However, their method included only physical steps in which observable control inputs were made (and did not include look, reach, or cognitive steps), and this may account, at least in part, for the difference in findings. The results of the current method in columns 1 and 2 would indicate that for the model implemented here, adding timing information is required if the models are to be of use. The results of correlating activity times and their ranks to real world data can be seen in the columns labeled 3 and 4. Although this shows a dramatic improvement over the simple step counts, could it be improved by adding the Modified MRT steps into the modeling methodology? By this addition the metric of Dual Task Conflict Potential (DTCP) is produced. This metric and its rank order are correlated in columns 5 and 6, and an activity time weighted version is shown in columns 7 and 8.

In general, correlations are higher between the modeling metrics and on-road data. This may be due to greater agreement in modeling the more conventional tasks that were tested in the on-road venue. The more complex tasks were performed only on the test track and these are also the tasks in which more variability between modelers was observed. General observations on each of these sets of metrics will be explored below. A more detailed summary of the analysis can be found in later sections of this chapter.

Columns 1 and 2 in Table 6-8 and Table 6-9 show the correlations for mean number of steps in a modeled task across all modelers (column 1) and the rank of that metric (column 2). The Step Count metrics have rather high correlation to both Static Task Completion Time and Total Shutter Open Time obtained from the Occlusion method. This is very reasonable, as the models were constructed with the mindset of modeling the task by including only the minimum activities required for the task while ignoring the activities required to drive the vehicle. Both of these Surrogates account for the time required to perform the task only. Step Count is also highly correlated with the DTCP metrics due to the calculation of DTCP. Step Counts, however, are rather variable between tasks and modelers and show no other significant correlations with the other metrics, either from the laboratory or driving environments.

Mean Total Activity Time generated from the Task Steps Assessment Models is well correlated with a number of metrics as can be seen in the columns 3 and 4 of Table 6-8 and Table 6-9. Correlations to time-related metrics include TSOT, Static Time, On-Road Task Duration, and STISIM Task Completion Times. This metric is highly correlated to On-Road Task Duration while being slightly less correlated to Test Track Task Duration. This indicates that the simpler tasks that were performed on-road are easier to model than the more variable tasks that were added for the test track. This high correlation, combined with the lowest correlation of the group being for the Static Task Completion Times, indicates that the models are representative of something in between single task performance and completing the task while driving.

Mean Total Activity Time has correlation with PDT Reaction Times, and the graphs exhibit a trend to group shorter, visual-manual tasks together then spread out the longer auditory-vocal tasks along a negatively sloping line. This grouping is similar to, but not as dramatic as, some of the patterns of separation that were seen with other correlation analyses. Interestingly this metric does not correlate well with the in vehicle OED detection reaction time metrics. The opposite is true however for detection rate metrics, with low correlation to PDT detection rate but higher correlation to the on-road and test track CHMSL detection metrics.

Mean Total Activity Time shows significant correlations to on-road longitudinal control measures such as Speed Difference, but not as well to the corresponding test track measures. In general, due to the larger number and variety of tasks performed on the test track, larger variations are found between tasks for all measures thus lowering correlations when compared with on-road performances. This difference was seen previously in the correlation analysis performed on laboratory surrogates as well. Rank ordering, however, improves the existing correlations slightly. Similar results are seen with lateral control measures but with fewer significant correlations. Large differences in how certain tasks were modeled by the individual modelers, however, tend to make these test track correlations lower than their on-road counterparts.

Mean Total Activity Time metrics also show numerous correlations to eyeglance metrics. High correlations exist for overall task glance measures as well as measures of glance behavior to specific categories of glances. There is good correlation to glances both to and away from the roadway. This is not surprising as tasks were modeled with the assumption that the task performer was driving a car but without any attempt to model the driving task. Surprising, however, is the lack of correlation to task-related glances for either venue. This effect may be explained with the simple fact that most auditory-vocal tasks have practically no task-related glances. This would imply that further examination of this metric might benefit from visual-manual versus auditory-vocal task division. This task type analysis orientation will be examined in the detailed sections to follow.

Columns 5 and 6 of Table 6-8 and Table 6-9 show no significant correlations for DTCP or the rank order of this metric to any of the other performance measures. With these correlations being so dramatically lower than the other metrics, one is forced to ask why, given the other

correlations here and the previous success of Modified MRT on which this measure is based. While the implementation of MRT may be at fault, a check of consistency within all individual's models revealed possible inconsistencies in the assignment of the MRT demand vectors for individual steps. Not all steps with the same Activity Verbs will require the same resources every time. However, it can be easy to input a demand vector for a step that is inconsistent with a similar step used previously in the model or in a model for a different but similar task. Corrections for inconsistency or changes in the calculation of DTCP may improve these correlations.

Table 6-8. Mean Analytic Surrogate Metric Correlations to On-Road Data

Metric	Road Data							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1 - Step Count	1.000							
2 - Step Count Rank	0.935	1.000						
3 - Activity Time	0.202	0.085	1.000					
4 - Activity Time Rank	0.237	0.152	0.935	1.000				
5 - Total DTCP	0.883	0.803	-0.087	-0.021	1.000			
6 - Total DTCP Rank	0.892	0.829	-0.010	0.064	0.981	1.000		
7 - Activity Time*DTCP	0.091	-0.023	0.970	0.903	-0.144	-0.068	1.000	
8 - Activity Time*DTCP Rank	0.195	0.067	0.930	0.968	0.008	0.093	0.924	1.000
Mean_TaskDur	0.266	0.197	0.932	0.898	-0.031	0.064	0.911	0.890
MdnTaskDur	0.264	0.195	0.936	0.896	-0.031	0.064	0.916	0.891
Mean_SDLP	0.231	0.218	0.658	0.757	0.099	0.160	0.678	0.748
Mdn_SDLP	0.254	0.215	0.668	0.765	0.153	0.200	0.690	0.764
Mean_SpeedDiff	0.276	0.206	0.908	0.907	0.016	0.093	0.891	0.893
Mdn_SpeedDiff	0.254	0.187	0.897	0.902	0.003	0.080	0.886	0.893
PctLVDecelMissRate	-0.186	-0.196	-0.610	-0.704	0.029	-0.031	-0.586	-0.675
PctCHMSL_MissRate	-0.083	0.013	-0.837	-0.824	0.209	0.167	-0.795	-0.819
MeanTskglncs	0.303	0.264	0.853	0.830	0.067	0.159	0.843	0.814
MeanTaskdur	0.291	0.219	0.932	0.897	-0.003	0.092	0.910	0.886
MeanTglsprs	-0.098	-0.005	-0.820	-0.875	0.182	0.099	-0.781	-0.870
MeanglncesRD	0.303	0.263	0.859	0.834	0.062	0.155	0.848	0.819
MeanduratRD	0.280	0.203	0.935	0.900	-0.021	0.075	0.912	0.891
MeanmeanRDdr	0.204	0.079	0.845	0.868	-0.102	-0.039	0.821	0.859
MeanmedRDdur	0.113	-0.007	0.831	0.850	-0.192	-0.125	0.822	0.851
MeangrateRD	-0.106	-0.015	-0.819	-0.880	0.171	0.090	-0.782	-0.873
MeanpctdurRD	0.067	-0.031	0.776	0.842	-0.183	-0.102	0.741	0.849
MeanglncesSA	0.222	0.163	0.926	0.867	-0.070	0.031	0.903	0.864
MeanduratSA	0.217	0.157	0.934	0.870	-0.073	0.028	0.914	0.869
MeanmeanSAdr	0.231	0.200	0.858	0.923	0.021	0.146	0.839	0.903
MeanmedSAdur	0.224	0.193	0.837	0.912	0.015	0.142	0.815	0.886
MeanglncesMR	0.223	0.167	0.922	0.865	-0.072	0.029	0.897	0.861
MeanduratMR	0.220	0.163	0.929	0.868	-0.071	0.029	0.907	0.866
MeanmeanMRdr	0.245	0.215	0.842	0.921	0.047	0.164	0.821	0.906

Metric	Road Data							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
MeanglncesNR	0.328	0.288	0.841	0.821	0.097	0.186	0.832	0.804
MeanduratNR	0.376	0.349	0.775	0.744	0.177	0.252	0.767	0.717
MeanmeanNRdr	-0.012	0.063	-0.655	-0.755	0.200	0.138	-0.630	-0.764
MeansdNRdur	-0.024	0.084	-0.737	-0.806	0.199	0.147	-0.718	-0.818
MeangrateNR	-0.081	0.012	-0.829	-0.877	0.205	0.120	-0.789	-0.873
MeanpctdurNR	-0.058	0.035	-0.779	-0.846	0.195	0.114	-0.745	-0.851
MaxTdur	0.306	0.245	0.797	0.825	0.013	0.112	0.798	0.803
MaxRDdur	0.306	0.245	0.797	0.825	0.013	0.112	0.798	0.803

Table 6-9. Mean Analytic Surrogate Metric Correlations to Laboratory Data

Metric	Lab Data							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1 - Step Count	1.000							
2 - Step Count Rank	0.886	1.000						
3 - Activity Time	0.271	0.152	1.000					
4 - Activity Time Rank	0.292	0.182	0.928	1.000				
5 - Total DTCP	0.898	0.896	0.085	0.092	1.000			
6 - Total DTCP Rank	0.844	0.924	0.065	0.080	0.973	1.000		
7 - Activity Time*DTCP	0.185	0.066	0.968	0.889	0.033	0.014	1.000	
8 - Activity Time*DTCP Rank	0.304	0.198	0.913	0.971	0.164	0.153	0.905	1.000
Mean_StaticTime	0.846	0.592	0.873	0.691	0.596	0.493	0.793	0.625
Mdn_StaticTime	0.832	0.573	0.879	0.702	0.571	0.467	0.793	0.625
Mean_TSOT	0.887	0.597	0.939	0.730	0.605	0.481	0.871	0.677
Mdn_TSOT	0.878	0.588	0.947	0.740	0.593	0.468	0.875	0.680
Mdn_Sti_SpeedDiff	0.661	0.515	0.575	0.711	0.400	0.361	0.525	0.672
Mean_Sti_Duration	0.522	0.412	0.876	0.875	0.292	0.280	0.836	0.855
Mdn_Sti_Duration	0.440	0.376	0.895	0.883	0.232	0.241	0.860	0.864
Pdta_Mean_MeanDetectRT	0.257	0.286	-0.712	-0.674	0.441	0.431	-0.695	-0.607
Pdts_Mean_MeanDetectRateRT	0.098	0.122	-0.844	-0.814	0.269	0.247	-0.844	-0.807
Pdts_Mdn_MeanDetectRateRT	0.196	0.205	-0.802	-0.764	0.365	0.334	-0.800	-0.741
Pdta_Mean_MdnDetectRT	0.133	0.166	-0.758	-0.747	0.333	0.323	-0.726	-0.675
Pdta_Mdn_MdnDetectRT	0.173	0.198	-0.718	-0.730	0.395	0.382	-0.679	-0.634
Pdts_Mean_MdnDetectRT	-0.013	0.023	-0.868	-0.856	0.184	0.165	-0.857	-0.843
Pdts_Mdn_MdnDetectRT	0.031	0.051	-0.847	-0.843	0.216	0.190	-0.833	-0.824
Strn_Mean_MdnRTCor	0.097	0.006	-0.730	-0.626	0.203	0.135	-0.701	-0.614
Strn_Mean_MdnRTAll	0.061	0.001	-0.726	-0.632	0.177	0.129	-0.705	-0.633

Columns 7 and 8 in Table 6-8 and Table 6-9 show correlations for DTCP multiplied by the Activity Time. Here it can be seen that the DTCP has no significant correlation to the data, and

that weighting it by Activity Time improved correlation to real-world task performance times, but decreases correlations to other measures when compared to Activity Time correlations. This would seem to indicate a potential value to completing the Modified MRT portion of this modeling method but shows that there are changes needed before the DTCP measures are a better metric than the Activity Times.

6.3.3 Summary of Modeling Findings

The Modified MRT Total Interference Potential models created at a task level for the DWM tasks show correlation to some laboratory and vehicle measures. These correlations were mainly for only the visual-manual tasks and largely caused by the outlier Destination Entry Task. These models also do not discriminate well among the tasks used here hindering discriminability of high and low workload as well as predictive power. Better results have been obtained for modeling driving performance, notably by Wickens and Horrey (1999). This however often involves considerable fitting of the model to a set of driving performance data. The goal of this modeling is to provide a tool that is useful for design prior to the availability of driving data and thus this method, while suitable in many applications, does not fully fulfill the requirements for this project's task modeling.

The Task Steps Assessment methodology, while much more complex, produces models that more accurately discriminate between tasks and have more correlations to both laboratory surrogate and real-world driving performance data. These models also yield a number of metrics, which can be used to examine tasks, the best of which currently is Activity Time. By analyzing at the task step level there is also greater potential for accurately representing tasks of a more complex nature. Thus this method meets the flexibility and predictive requirements of this project. The system is however by no means perfect, there are issues to address further.

In analyzing the Task Steps Assessment models, a number of differences can be seen both between and within individuals. Due to variability between modeler's styles, rank orders are typically correlated more highly on an individual basis than are the raw scores for any of the metrics from the Task Steps Assessment. Higher correlations can be seen however when using the mean values of a group of individuals' models. This can, in part, be accounted for by the differences in each individual's modeling styles. These differences are called Modeler Strategy and include the individual's understanding of the methodology and the task, the level of detail to model to, differences in use of the by-pass step determine, and various initial and goal state differences.

Another cause for variation is the Task Strategy or method a modeler believes a person will use to perform the task. Two dramatic examples are the Sports Broadcast task and the Book-on-Tape task. In Sports Broadcast, three of four modelers assumed a person would listen for a keyword without processing any information not related to that keyword. The fourth modeler assumed that a subject had to process each portion of the recording to determine if it was the information that they had been asked to remember. In Book-on-Tape, three of four modelers created very simple models of listening and retaining information. The remaining modeler broke the story into a number of key elements, processing and retaining each one. This resulted in a vast difference in both Task Step Count and Activity Times for both Book-on-Tape tasks.

The Dual Task Conflict Potential metrics currently do not improve model performance. As noted previously, this may be due in part to inconsistency in assigning resource demand vector elements. Improvement to the modeling utilities may aid in consistency and improve correlations. Alternately, allowing for levels of a resource demanded may improve performance. Currently demand vector elements are binary values. Thus all task steps with a particular resource demand are assigned the same value for that demand. Allowing for levels of demand for each of the vector elements may help to differentiate task steps more and thus improve correlations.

6.4 Discussion of Task Steps Assessment Results

In review of the task analysis results, a number of areas of potential improvement are apparent. The first area is modeler training. Examples and instructions have been updated with clarifications to help eliminate differences in understanding. This should aid consistency between modelers, however, there still exist strategy differences to be corrected.

Modeler strategy is the way in which an individual applies the system to the modeling of a task. Inconsistency among modelers arises due to factors such as inclusion or exclusion of task steps, level of detail in steps that are modeled, initial and goal states, and application of the activity verb vocabulary. Thorough training of modelers, with a background in a related field such as Human Factors or Cognitive Psychology, in the application of the Task Steps Inventory methodology should improve inter-modeler consistency.

Task strategy is the method that a modeler believes the subject will use to complete a task. In modeling complex real-world tasks this type of difference between modeler's cannot be overcome. People will usually perform tasks in a manner that best suits their own traits and skills. A method for analyzing models to identify where differences in individual's models are due to different Task Strategies could be useful, especially to illuminate possible improvements in prototype designs.

When multiple modelers work on a task set, the two strategy differences (modeler and task) can become more apparent. Mediation of the models by the group of modelers could be implemented to construct a final model or models. In this way, modeler strategy differences can be eliminated and task strategy differences can be identified.

If the Modified MRT component is to be used with this modeling methodology, improvement is needed in the calculation or weighting of the DTCP. Correlations of DTCP to the other data are not significant unless they are weighted with Total Activity Time. In most instances, this weighted measure is less significantly correlated with the other data than is Activity Time. This would indicate that the DTCP is not adding a valuable metric to the output of the methodology and change is needed.

Another potential improvement to the DTCP generation in the Task Steps Inventory is to provide a default demand vector in the spreadsheet tool for each Activity Verb. When modeling, it can be difficult for an individual to remain consistent between steps in a model and between models when entering the demand vector. A generally agreed upon default demand vector could be added to the spreadsheet tool for each Activity Verb. The modeler could then change the default values for each resource based on the particular characteristics of the individual step, thereby eliminating a large area of potential inconsistency from the modeling methodology.

6.4.1 Count-Task Steps Analysis

The first metric that is obtained from the Task Steps Assessment methodology is the sum of the number of physical and cognitive steps in the model. This measure is well correlated with Total Shutter Open Time and Static Task Completion Time as was shown in the previous section. This section will examine the Step Count metric among modelers as well as Mean Step Counts in more detail.

The first comparison to be made is between modelers with the raw Step Counts. In Figure 6-3, a comparison is made between the two modelers with the most experience with subjects performing tasks. The graph shows a tight clustering of tasks along the diagonal. In addition, these modelers were the first to apply the methodology. Given these facts, good agreement should be expected. However, the HVAC and Destination Entry tasks stand out from the rest of the tasks. For HVAC, this is explained by a Task Strategy difference, one modeler has fewer steps due to having only

one verification step at the end of the subject setting the control knobs. The same is true for the Destination Entry Task, where one modeler assumed that an expert user would scroll up a menu, the shorter way to get to the target, thus requiring fewer task steps.

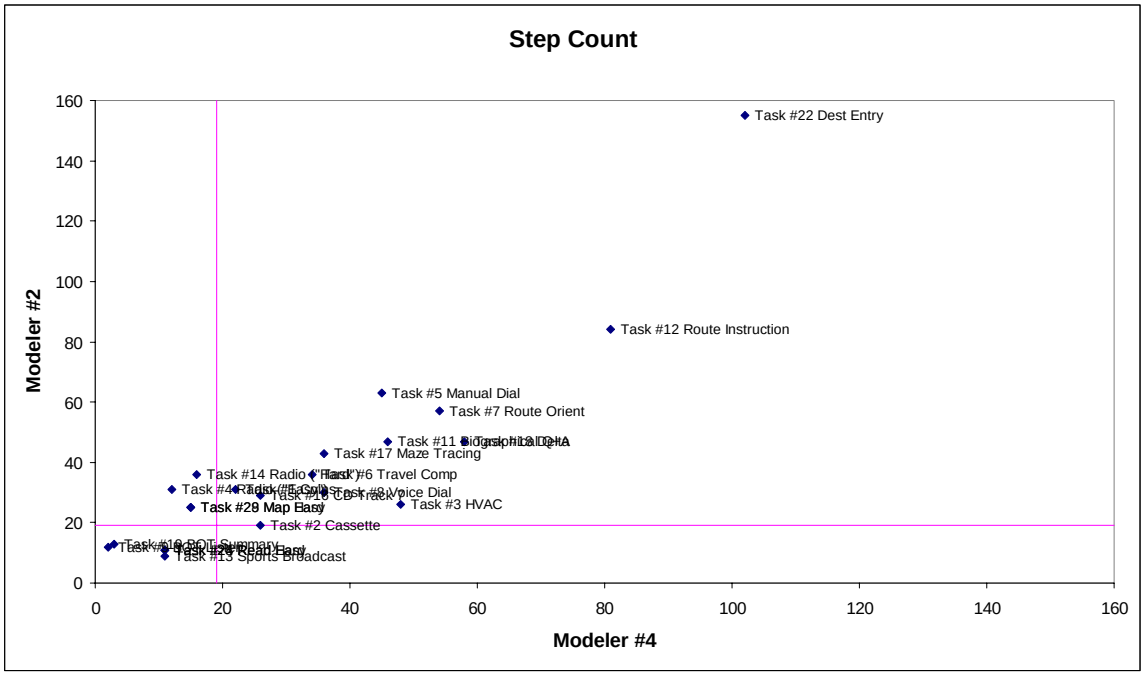


Figure 6-3. Step Counts for Modeler #2 Versus Modeler #4

Figure 6-3 indicates a Modeling Strategy difference between the two modelers with a couple of tasks. The most obvious of these is the Destination Entry task, while both individuals have it near or at the top in number of steps, the difference in steps is very large. This is because one modeler applied a Visual Search step and one button press to scroll through each screen of the scrolling list navigation unit. The other modeler applied a Compare step for the first entry in each menu and then button presses to get to the next screen. In this instance, the Visual Search step is being used in place of numerous physical and cognitive steps. Route Instruction in the same figure represents another type of modeling strategy difference, that of level of detail. Here, one modeler used the combination of a Listen and Speak step pair for each passage whereas the other used a more detailed Listen and Retain for each detail of a passage, followed by combinations of Recall and Speak for each detail in the repeat phase of the task. Manual Dial stands out in the group of tasks as well. This again is similar to Destination Entry in that a single button-press step was used by one modeler, while the other more accurately used the same step for each digit of the phone number to be dialed.

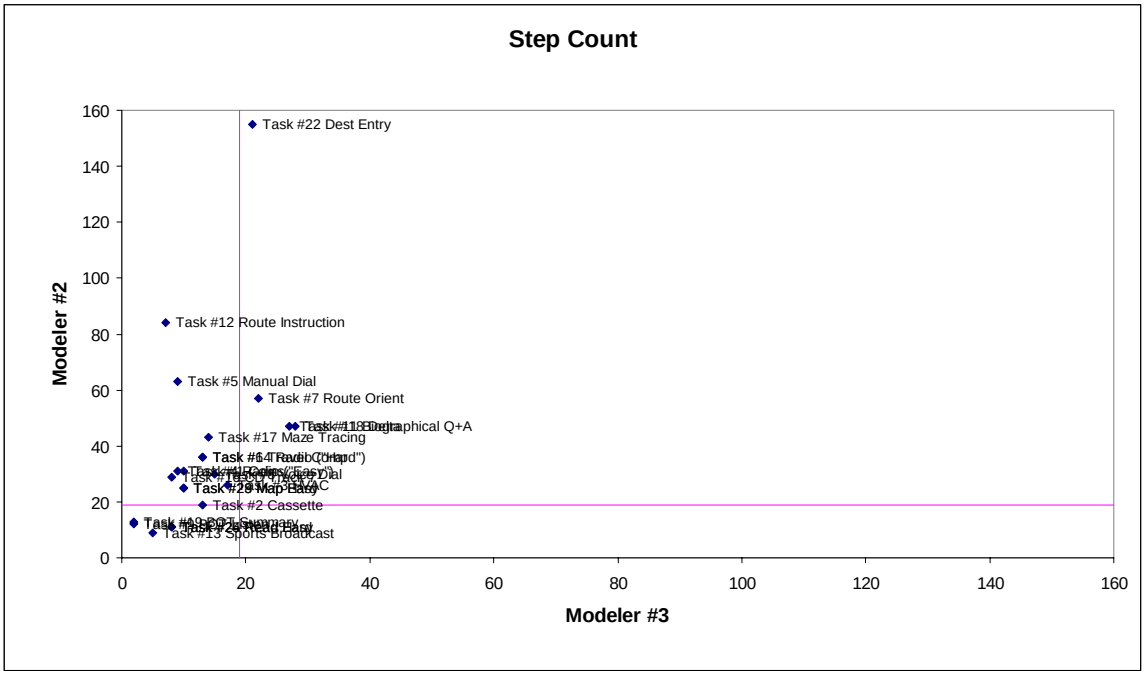


Figure 6-4. Step Counts for Modeler #2 Versus Modeler #3

As can be seen in Figure 6-3 and Figure 6-4, due to the variety in number of steps in individual models, the graphs can be difficult to read. Rank ordering the tasks based on the number of steps is a way to improve the readability of these graphs.

In Figure 6-5, the Rank Order of Step Count for two modelers is used with vertical and horizontal reference lines to help distinguish potential model differences. With these reference lines, the graph is divided into quadrants. With consistent modelers, one would expect to see tasks located along a diagonal from the lower left to upper right. This graph exhibits this diagonal tendency with a couple of exceptions. Manual Dial stands out with one modeler ranking it far above median and one far below the median. This is due to the same button press issue that was previously detailed for another individual's model. Radio (Easy) also stands out somewhat in this graph. This is due to a level of detail difference as was previously discussed. The last outlier is Sports Broadcast, this again represents a Task Strategy difference. One modeler assumed a subject would listen to and process each of the details of the broadcast while the other assumed they would listen for a keyword (the team in question's name) and only process and retain that information.

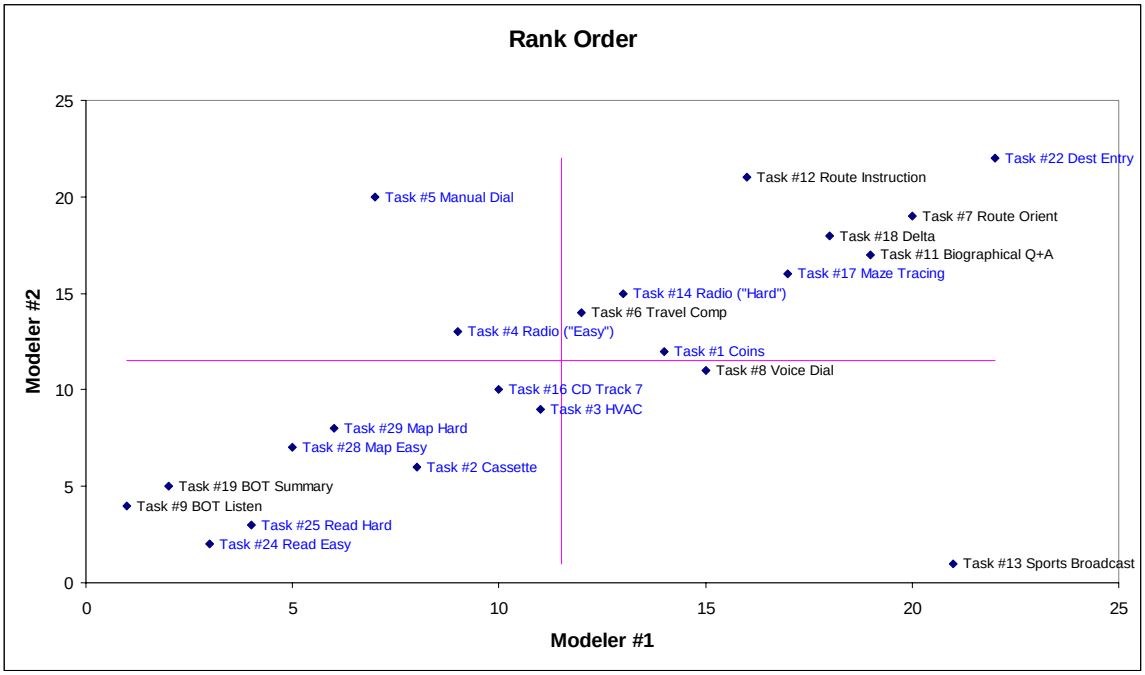


Figure 6-5. Rank of Step Counts for Modeler #1 Versus Modeler #2

Figure 6-6 is an example of the most disagreement and shows many points with the types of differences that are detailed above. Still this graph shows that the Rank Order of Step Counts between modelers reliably places most tasks in the expected quadrants and in relatively the same locations within the quadrants.

Figure 6-6. Rank of Step Counts for Modeler #2 Versus Modeler #3

When the group of models is compared as a whole in Figure 6-7, one can examine the trend in step counts both within and between modelers. However, it is difficult to resolve relative similarities due to modeling style differences between individuals. Figure 6-8 shows the Rank Order of Step Counts for all modelers. This graph has the same characteristics as the Step Count version but shows more agreement in task ratings. The use of rank ordering shows modelers' agreement more clearly by eliminating some of the consistent differences in modeling style between modelers.

While individual sets of models typically show lower correlations with the other metrics, the medians of the group of models show better correlations. This is due to the fact that with only one set of models it is rather difficult to identify the Modeling and Task Strategy issues discussed above.

Figure 6-9 shows the correlation between Mean Total Shutter Open Time (TSOT) and Mean Step Count. This graph has a tight cluster of tasks in the lower-left quadrant with Destination Entry the sole point in the far upper-right quadrant. In this instance, Destination Entry is a leverage point. Due to the nature of the Destination Entry task, it is often reported as an outlying point and in this case it is responsible for the high correlation seen in section 6.3 of this chapter and the relatively high R^2 value seen in the graph. When Destination Entry is removed from the task set, as in Figure 6-10 the R^2 for this graph drops to an insignificant level and the graph shows no structure or clustering of tasks. The same behavior is seen with the other significant correlations for Step Counts with Mean Total Shutter Open Time and Mean and Median Static Task Completion Time.

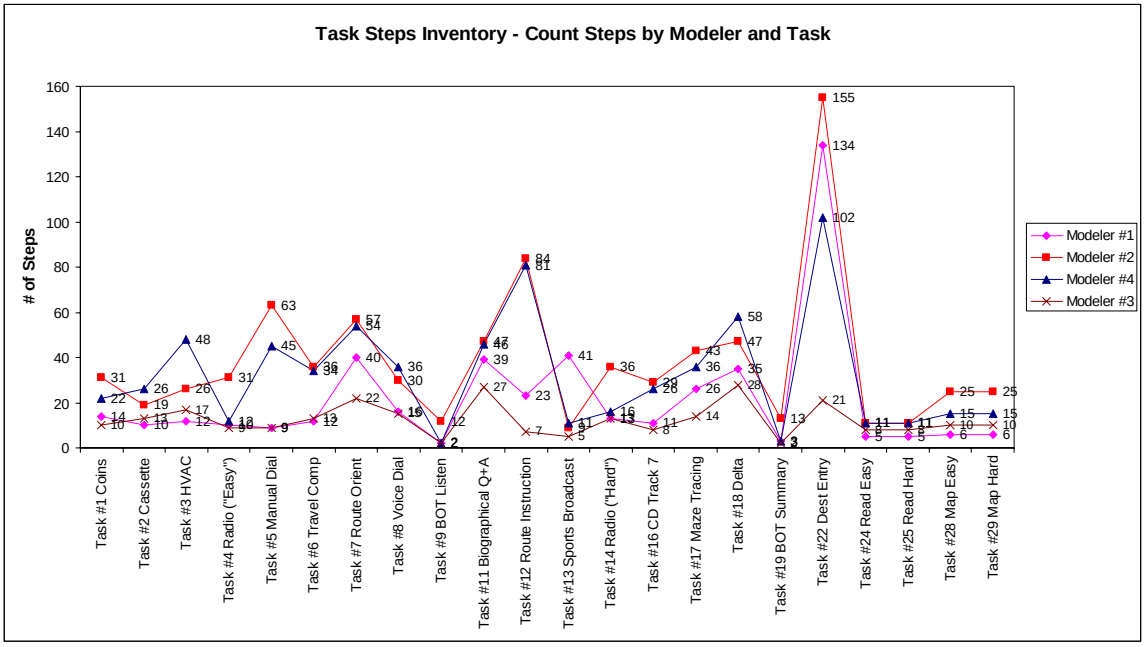


Figure 6-7. Step Counts for All Modelers

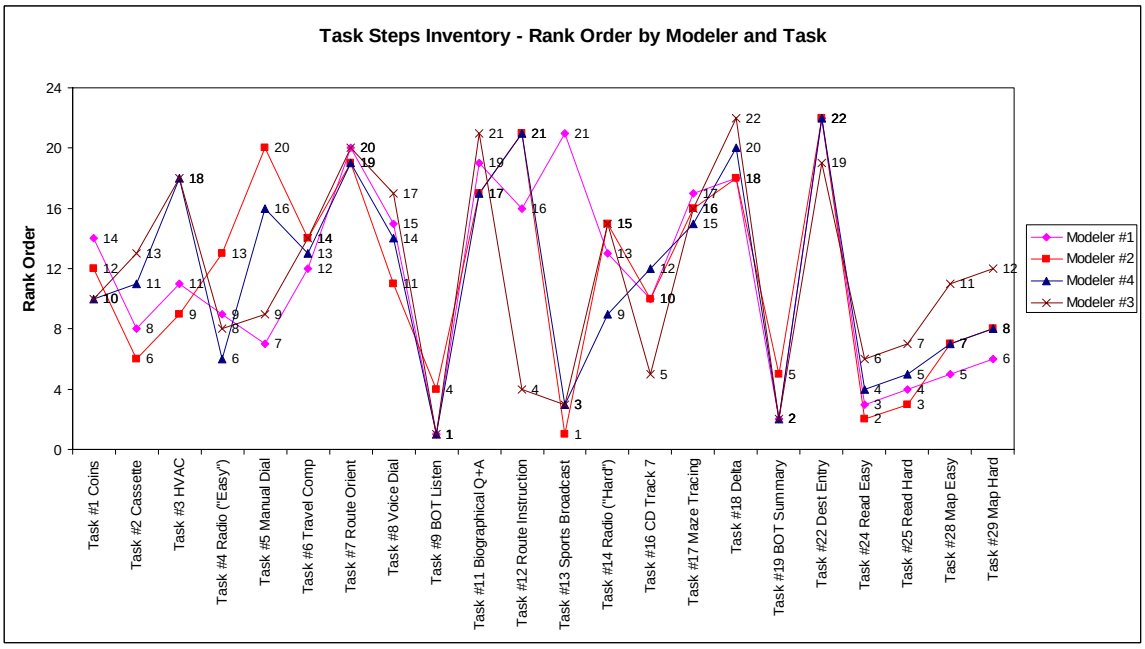


Figure 6-8. Step Count Ranks for All Modelers

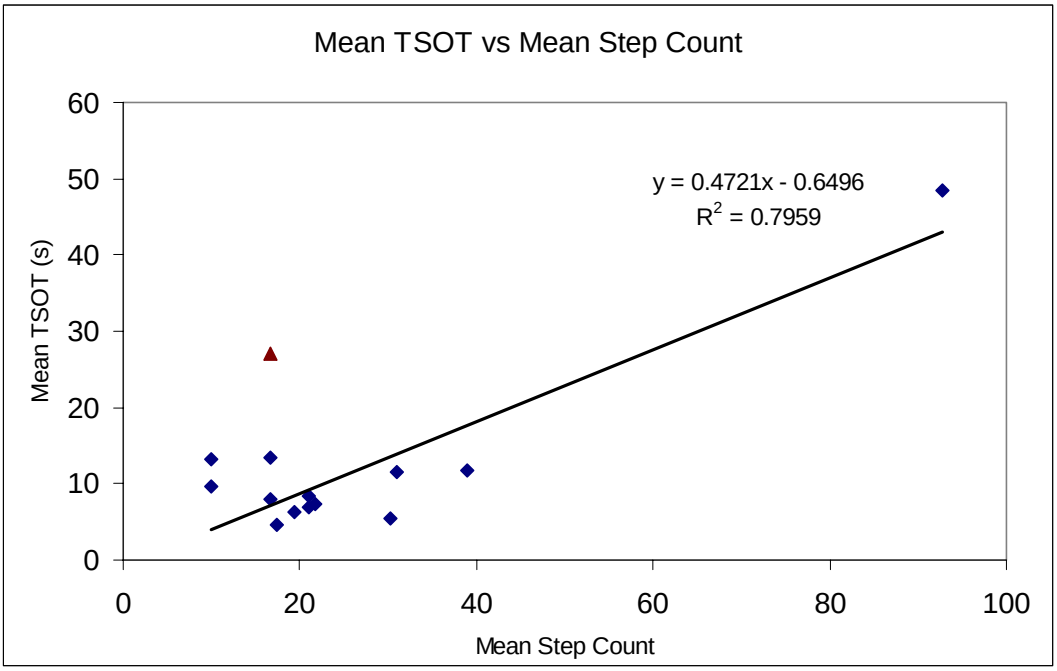


Figure 6-9. Mean Total Shutter Open Time (TSOT) Versus Mean Step Count

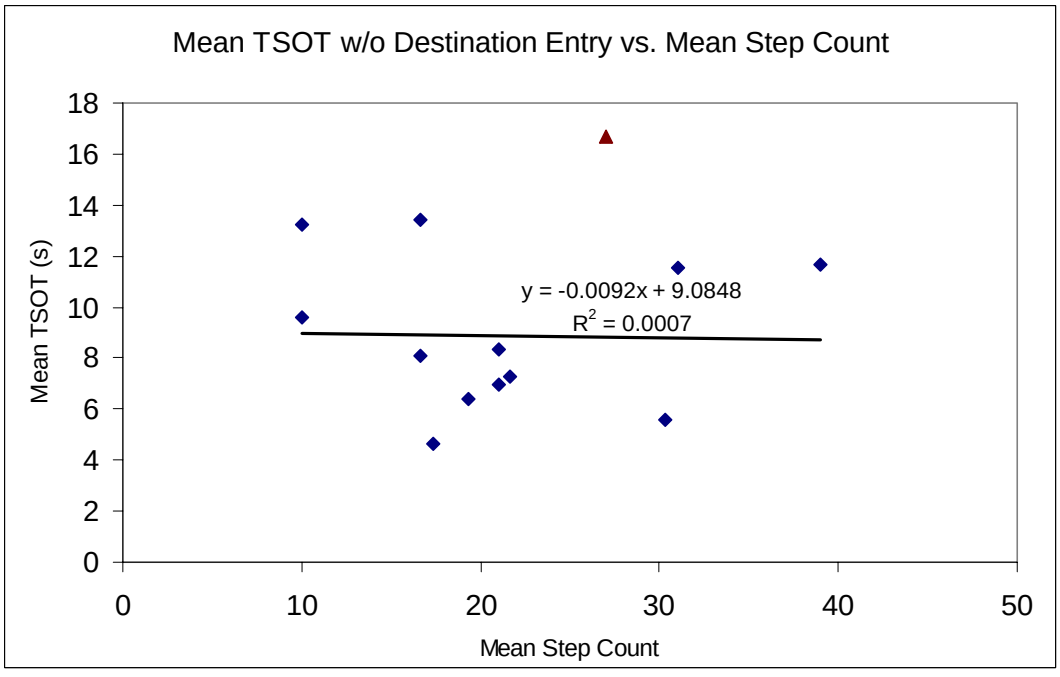


Figure 6-10. Mean Total Shutter Open Time (TSOT) without Destination Entry Versus Mean Step Count

6.4.2 Count-Task-Steps Results

Overall, the count of the number of steps in a Task Steps Assessment is not predictive of any of the other measures examined here for secondary task performance. This is due largely to the highly variable ways in which analysts modeled tasks. Step Counts can help to identify major Task Strategy differences, but due to Modeling Strategy differences the raw counts yield little insight into task performance. This can be improved slightly with Rank Ordering of the tasks. Still uncorrelated with other measures, rank orders do help to eliminate the effects of more systematic modeling strategy differences between modelers. This aids in identifying task strategy differences, which are important to find as modelers can essentially be examining different tasks if there is a large task strategy difference between them.

The most variability between modelers is seen with the more complex tasks involving more cognitive or "unobservable" steps. Training and experience with the methodology as well as a background in a related field can help to reduce these differences. More study of cognitive processes could also offer guidance that may augment the modeling instructions and help to reduce variability.

Mediation by a group of modelers to determine a final model for a task may also help to improve correlations. Even if this is not the case, mediation should be performed so that the best possible model is used for input to the Activity Times portion of the methodology or to other modeling architectures.

6.4.3 Task Activity Time Analysis

The next metric that is obtained from the Task Steps Assessment methodology is the sum of the Activity Times for each physical and cognitive step in the model. This measure is well correlated with many measures related to task time, OED detection, and vehicle control for the laboratory, on-road, and test track venues as was shown in section 6.3. This section will examine in more detail the Activity Time metric between modelers as well as Median Activity Times.

As with Step Counts in the preceding section, Activity Times show differences between modelers. In Figure 6-11, agreement between two modelers is rather high with tight clusters of tasks around the diagonal, the exception being the Route Instruction task. This difference is a combination of modeling and task strategy differences between the two modelers. The modeling strategy differences are caused by differences in times assigned to Wait and Listen steps and the use/non-use of concurrency of steps. There is also a task strategy difference represented by the addition by one modeler of Compare steps, assuming the subject will compare with memory the verbal answer returned at the prompt to repeat the directions. While there is significant difference in Activity Times, the two models contain 81 and 84 steps. This significant difference in one measure but not the other seems to be an indication of modeling strategy differences and appears in other comparisons as well.

Figure 6-11 presents the quadrant graph version (presented in the last section as well) of the graph in Figure 6-12. In this graph, there are four tasks that appeared in the Activity Time graph as reasonably clustered with the other tasks, the Rank Order graph, however, has set them away from the cluster of other tasks. Three of these lie near the median rank line (given that value is 10.5 this is expected with these graphs), however, the fourth sits off the line. When these two Maze Tracing models are examined, significant modeling strategy differences are found. One modeler used the Search and Select cognitive steps to represent the process of determining a route through the maze while the other simply used Look and Move steps to represent both the cognitive and physical steps involved. This graph also shows good clustering of visual-manual and auditory-vocal tasks.

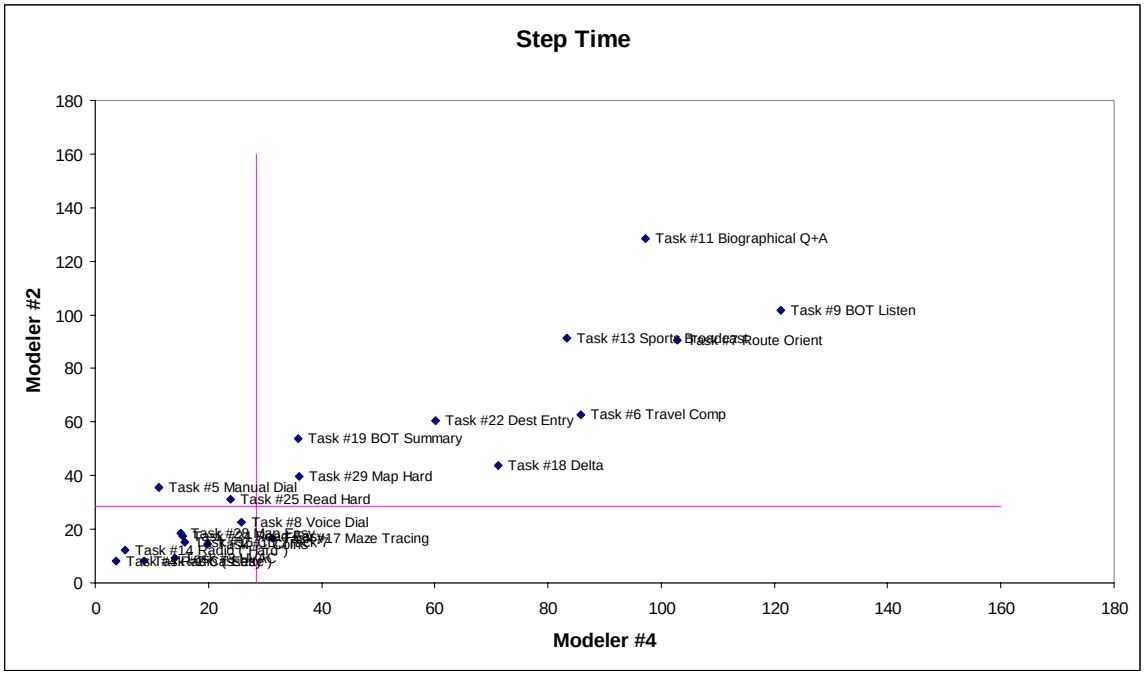


Figure 6-11. Total Activity Time for Modeler #2 Versus Modeler #4

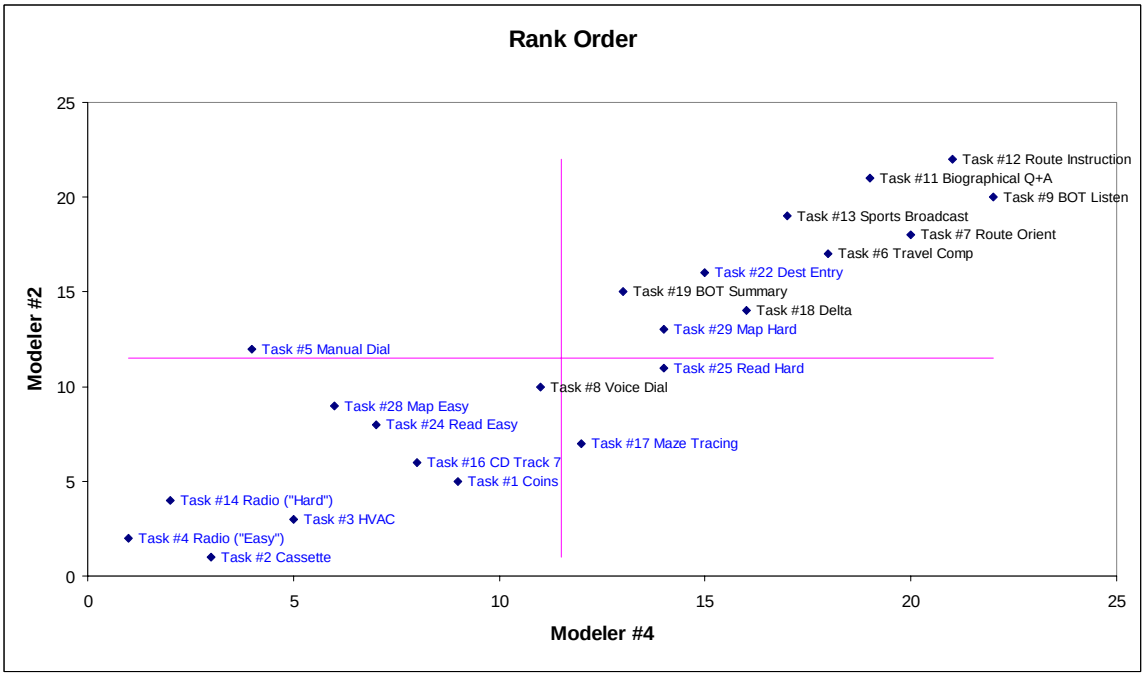


Figure 6-12. Modeler #2 Versus Modeler #4 Total Activity Time Ranks

Figure 6-13 shows the least agreement between modelers when assigning Ranks based on Activity Times. In this case, one modeler has drastically lower Activity Times than the other, however, by ranking the tasks they can be compared directly and some of the modeling strategy differences are eliminated from the picture.

Three tasks Read (Easy), Route Orientation, and Book-on-Tape Summarize do, however, still stand out, due to very large differences in modeling strategy. Modeler #3 typically used far few steps than Modeler #4. This difference is mostly eliminated by rank ordering. However, Read (Easy) stands out due to the trend of fewer steps mentioned above. So many other tasks have shorter Activity Times that even though these two models agree well in number of steps and Activity Times, their ranks are very different. With Book-on-Tape Summarize, one modeler assumed a very short paraphrasing of the story while the other assumed longer. This caused the two models to have the same number of steps but very different Activity Times and thus different ranks. The Route Orientation task stands out due to Modeling Strategy differences, mostly of the level of detail sort. One modeler did not include Retain and Wait steps, one modeler used these steps because the performer must remember the heading and adjust for incorrect answers to proceed to the next passage. Since the task is timed by the recording, the Wait steps were also necessary to represent the task of waiting for the next passage to begin.

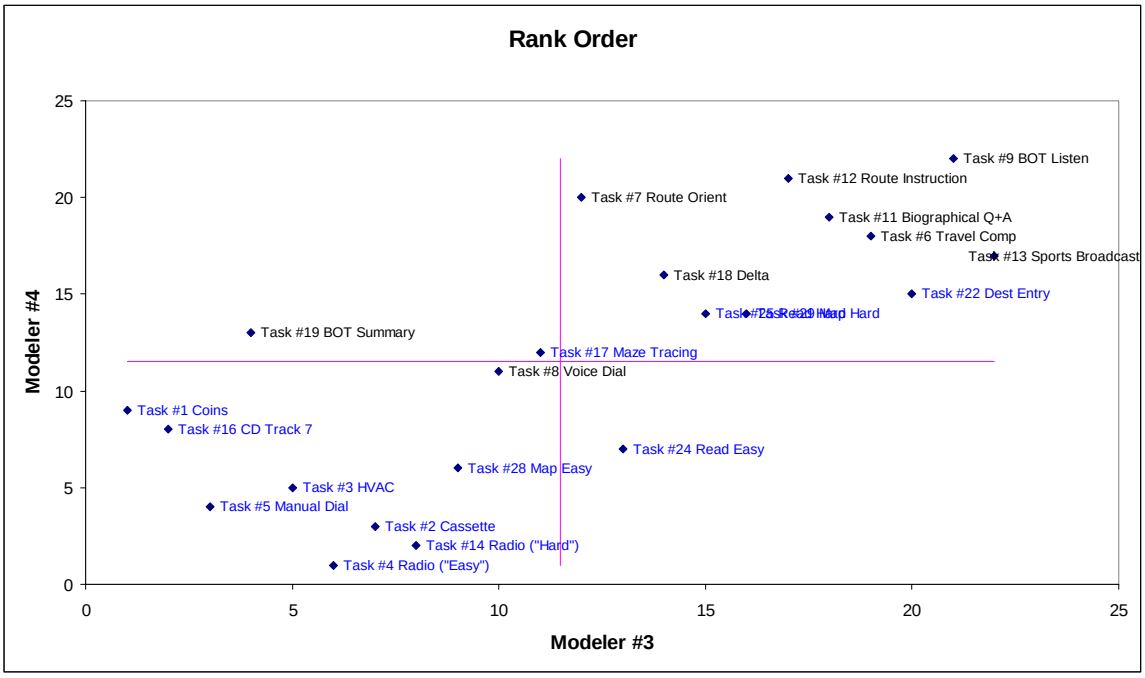


Figure 6-13. Modeler #3 Versus Modeler #4 Total Activity Time Ranks

These three graphs (Figure 6-11, Figure 6-12, and Figure 6-13) show both close and loose agreement between modelers based on Activity Times and Activity Time Ranks. The quadrant graphs show that rank ordering can remove systematic differences while still making drastic differences in modeling and task strategy stand out for further examination.

When the models are grouped together, trends can be examined and problems illuminated that may not necessarily show on other graphs. In Figure 6-14, it is easy to see agreement in groups of tasks on either end of the x-axis, but large differences are apparent for other tasks such as Route Orientation, Biographical Q&A, and especially Route Instruction, although it is difficult to

determine the cause. When the Activity Times are first ranked and then graphed as in Figure 6-15, a cleaner graph is produced and some of the systematic differences are eliminated. This leaves a clearer picture of which tasks have significant differences between models, such as Radio Easy and Book-on-Tape Summarize.

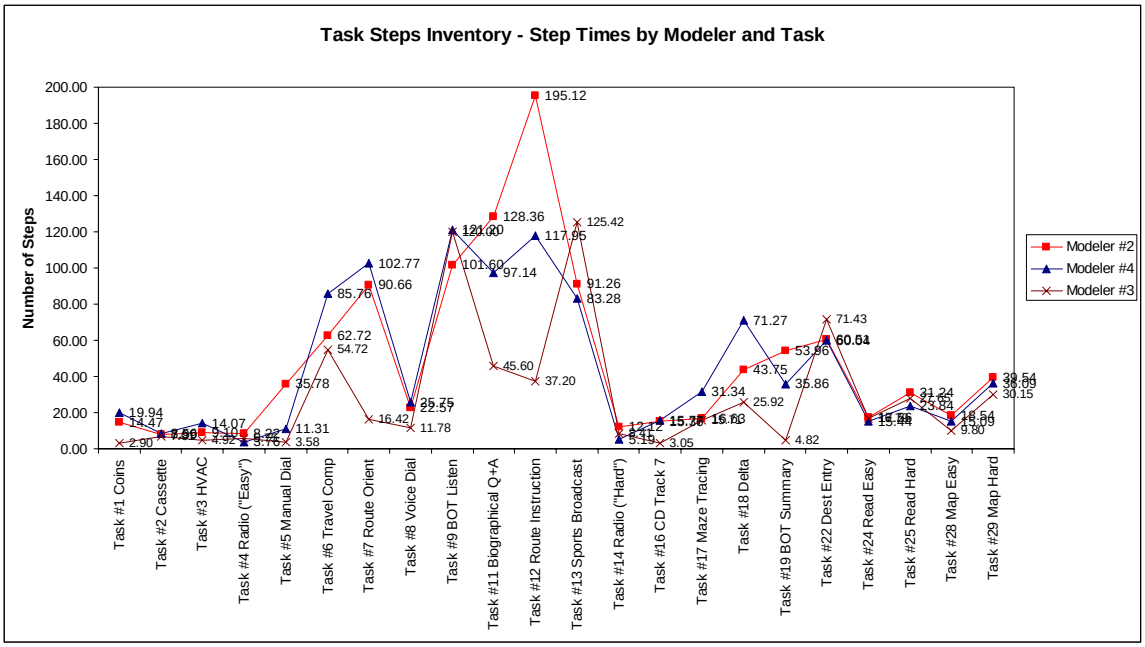


Figure 6-14. Total Activity Time by Modeler and Task

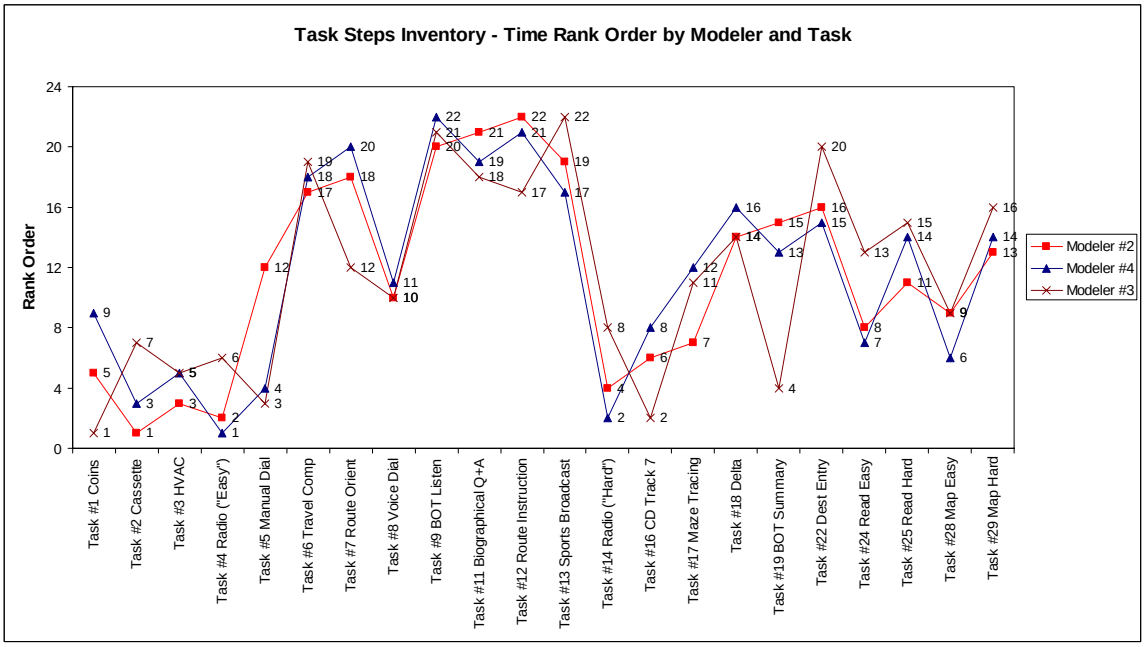


Figure 6-15. Total Activity Time Ranks by Modeler and Task

Table 6-8 presented earlier in this chapter shows high correlations between Median Total Activity Time and the surrogate, on-road, and test track data. While such a table is useful, plots of the measures in question tell a much more detailed story. In this section, graphs of some of the more relevant correlations will be presented and discussed.

The first comparisons to be made between Mean Total Activity Time and other measures are the task performance time measures. As was seen for the Step Count metric, there is a high correlation between Median Total Activity Time and both Median TSOT and Median Static Task Time. As was the case with Step Counts, when these metrics are plotted against each other, Destination Entry is a far outlier and when it is removed, the significance of R^2 drops below 0.50 so these graphs are not presented here.

The next logical time-to-time comparison would be Mean Total Activity Time versus the while driving task performance times. Figure 6-16 shows the correlation with On-Road Task Time. Here, the visual-manual tasks are (plotted against the secondary Y-axis on the right) grouped tightly along the regression line. The auditory-vocal tasks, which by their more cognitive nature are harder to model, lie in the upper right with much less clustering. Inter-modeler differences with Travel Computation, Route Orientation, and Route Instruction, as well as the fixed task duration, make these tasks difficult to model and lowers correlation to Task Time. A similar pattern can be seen in Figure 6-17 where Mean Total Activity Time is compared with Test Track Task Time. Here the visual-manual tasks are again clustered (except for Destination Entry) and the auditory-vocal tasks are more spread out. The R^2 values for both venues are rather high for the visual-manual tasks while the auditory-vocal tasks have values less than 0.5. As is seen with other metrics, Book-on-Tape-Summarize and Voice Dial tasks are grouped with the other short tasks while Delta Flightline (Delta Flight Information) and Destination Entry both are clustered with the longer fixed duration tasks.

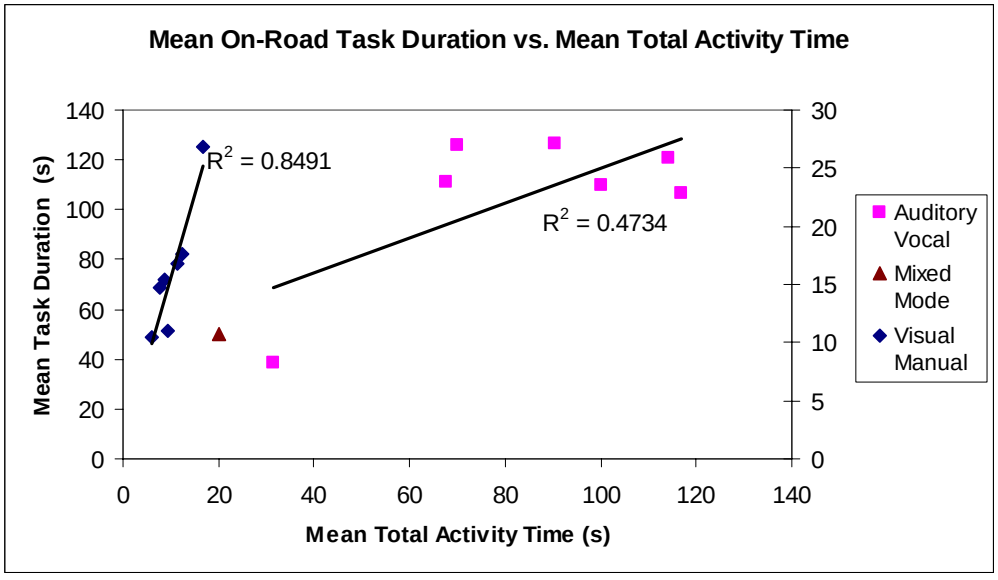


Figure 6-16. Mean On-Road Task Duration Versus Mean Total Activity Time

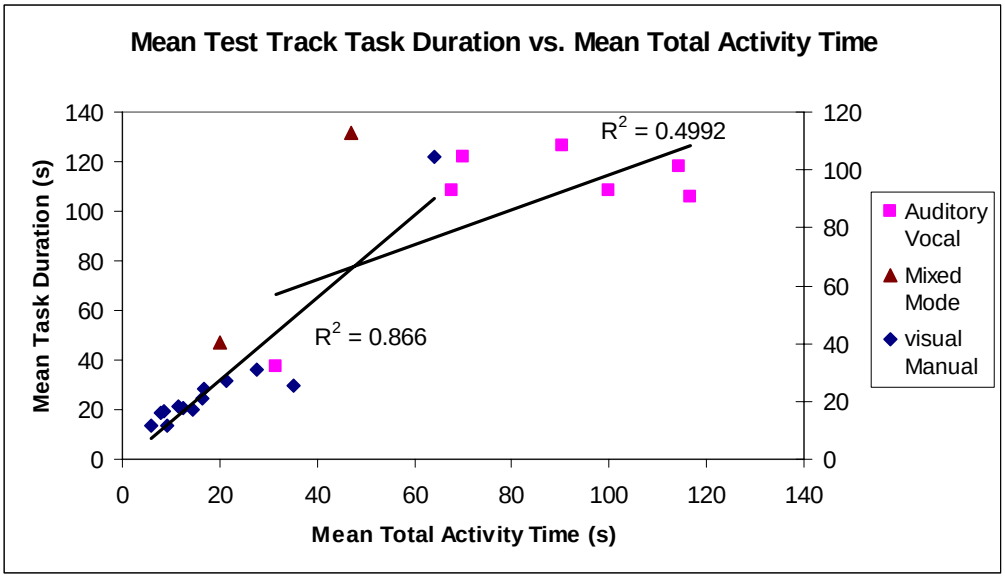


Figure 6-17. Mean Test Track Duration Versus Mean Total Activity Time

When Mean Total Activity Time is plotted against STISIM Task Time, a graph is produced which is nearly identical to the test track version. The R^2 of this graph is 0.87 for visual-manual tasks and the only distinct difference is that most tasks have a longer duration in STISIM than the test track. This difference may be due to the lack of "road feel" with the simulator and the high visual demand of these tasks.

An interesting group of correlations that appeared in Table 6-8 was that between Activity Time and Reaction Times, all of which were negative and several of which were at significant levels.

Figure 6-18 shows longer reaction times for Sternberg detection for tasks with longer total activity times. This is similar to the results of comparing other actual task performance times to the various reaction time measures. In short, Activity Time displays the same short task time/long reaction time behavior seen with other metrics.

Figure 6-19 shows the correlation for Total Activity Time versus PDT with STISIM Mean Median RT. This graph also shows, yet again, the clustering of the shorter visual-manual tasks separately from the longer auditory-vocal tasks.

For both of these comparisons the overall task set has reasonably high correlation for Total Activity Time and the surrogate metrics. When divided by task type however the correlation for the visual-manual task type becomes insignificant. The correlation to auditory-vocal task type remains however, this may indicate that for the auditory-vocal tasks Total Activity Time is a useful predictor for this task type.

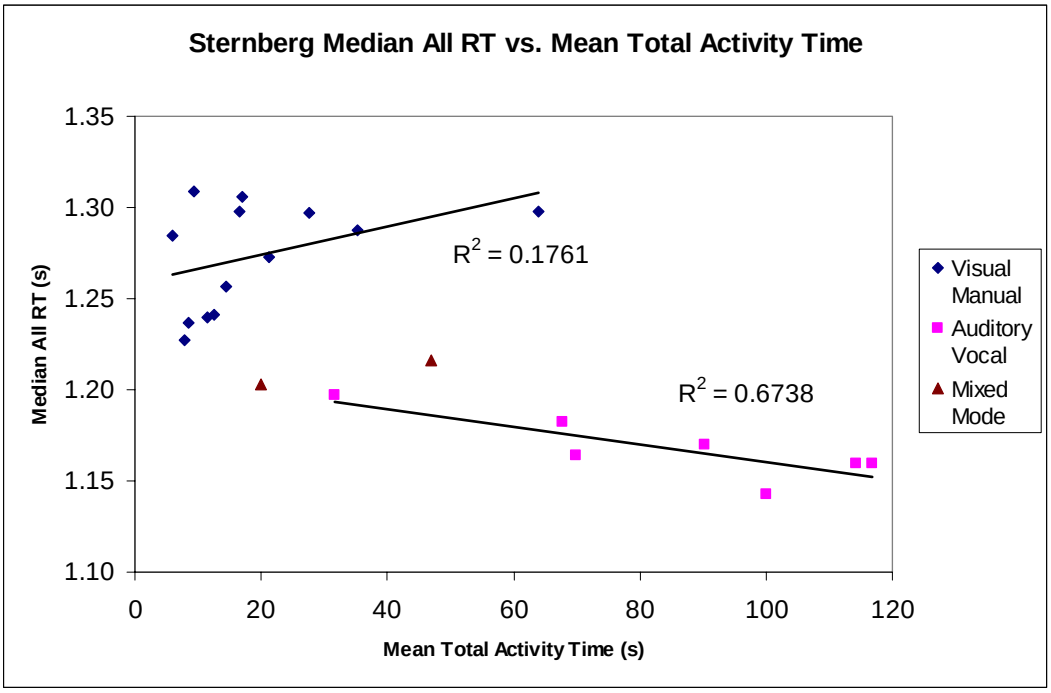


Figure 6-18. Sternberg Median All Response Time Versus Mean Total Activity Time

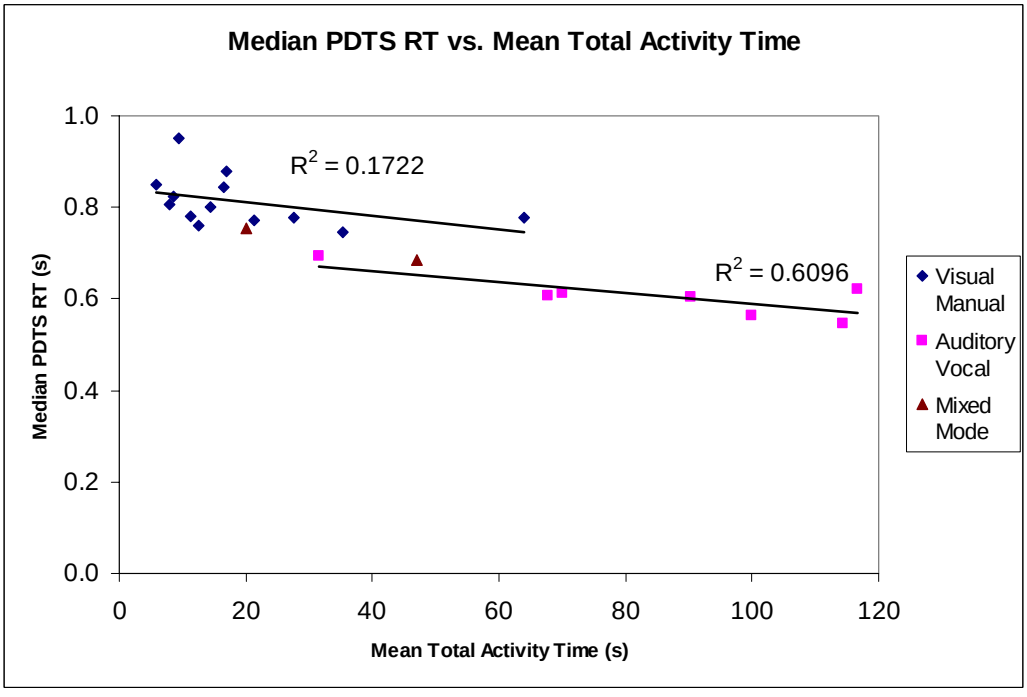


Figure 6-19. Median PDT with STISIM Response Time Versus Mean Total Activity Time

Table 6-8 also shows high negative correlations between Median Total Activity Time and the CHMSL Percent Miss metric for on-road data. Figure 6-20 shows this correlation for the On-Road measure.

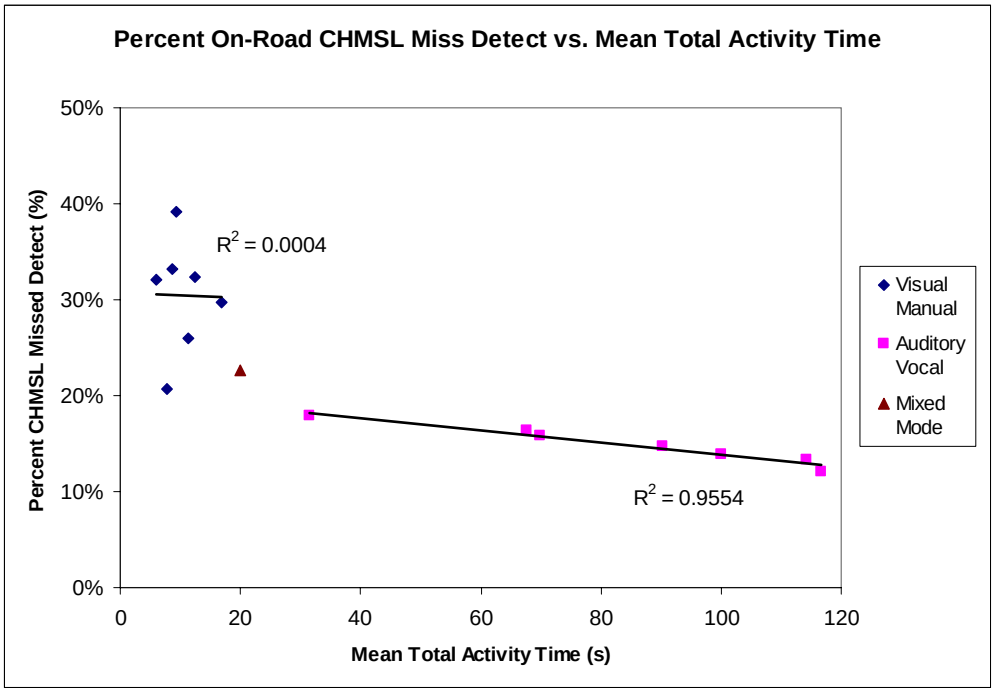


Figure 6-20. Percent On-Road CHMSL Missed Detections Versus Mean Total Activity Times

This graph, as with others, discriminates the shorter visual-manual tasks from the longer auditory-vocal tasks. For the overall task set, correlation is high, however when the tasks are divided by type, correlation again remains for only the auditory-vocal tasks.

So far the correlations that have been discussed have been related to task performance times and to OED detections. A number of correlations between Median Total Activity Time and vehicle control measures also showed significant correlations.

Figure 6-21 shows the Rank Order of tasks based on Mean Total Activity Time plotted against the On-Road Median SDLP. For this metric, correlation is only significant for the visual-manual task type. The task types, however, were all well segregated with the mixed-mode task separating the two groups of tasks.

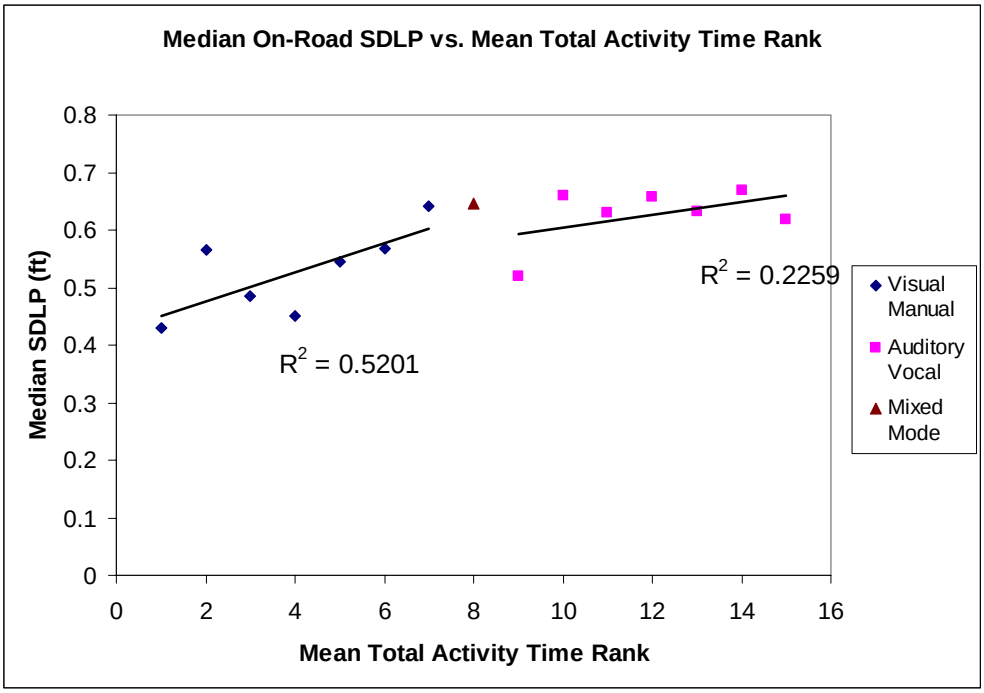


Figure 6-21. Median On-Road SDLP Versus Mean Total Activity Time Rank

For the entire task set, correlation between Mean Total Activity Time and Test Track SDLP is rather high, when the tasks are examined by type, correlation remains for both visual-manual and auditory-vocal tasks as can be seen in Figure 6-22. Figure 6-23 and Figure 6-24 show the correlation between Mean Total Activity Time and Speed Difference for the on-road and test track venues. Similar to the SDLP comparison, the high correlation for the entire task set is reduced when the task types are correlated separately. However in this comparison slightly better correlation is retained for the visual-manual tasks. Correlation for the visual-manual task set for the on-road venue places the tasks common to both venues similarly in both graphs. The improved correlation for the test track venue is in part due to the Destination Entry task being such an outlier. Without Destination Entry, correlation is not as high but it is still significant for the test track. As with other metric comparisons, some of the tasks that lie off the regression line exhibit task and modeling strategy differences that may be affecting the comparison.

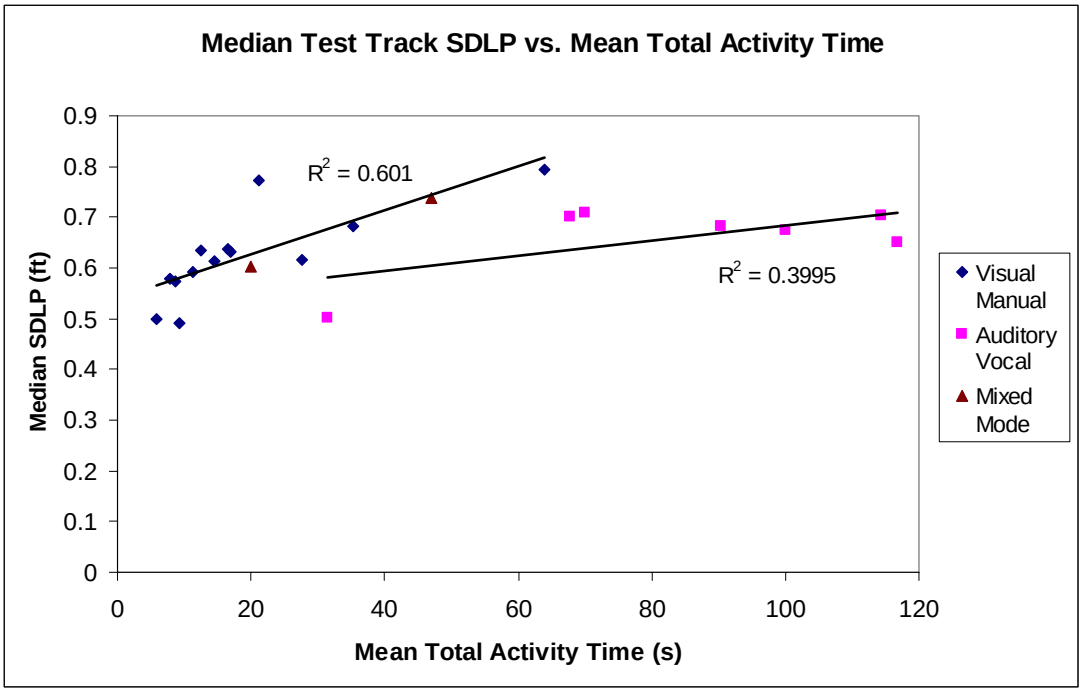


Figure 6-22. Median Test Track SDLP Versus Mean Total Activity Time

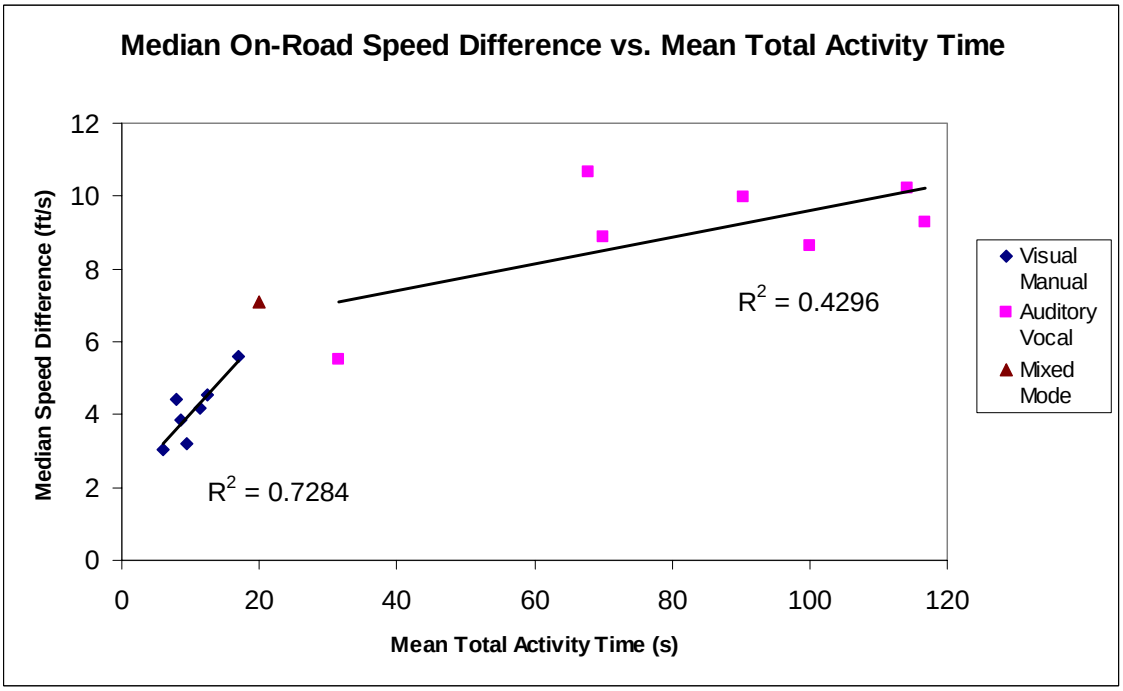


Figure 6-23. Median On-Road Speed Difference Versus Mean Total Activity Time

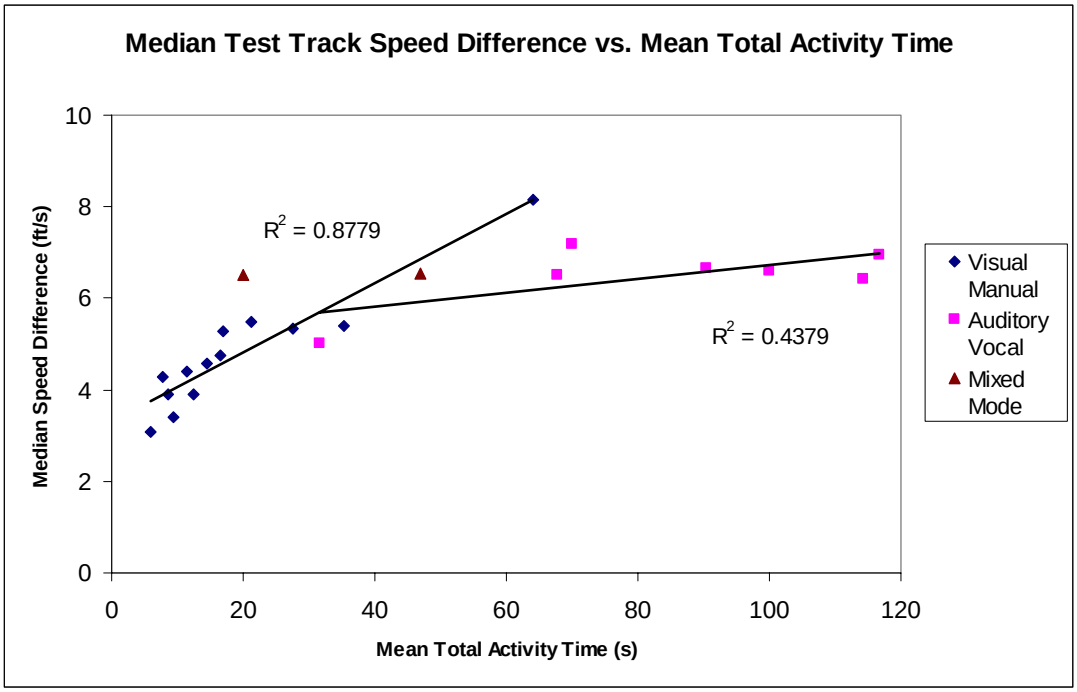


Figure 6-24. Median Test Track Speed Difference Versus Mean Total Activity Time

Total Activity Time has many significant correlations with eyeglance metrics. Significant task effects on the eyeglance metrics, closely related in some cases to task duration, heavily influenced Total Activity Time correlations. In order to examine these relationships, task types are correlated separately again and as before, correlations are lowered in general. Test track and on-road graphs are similar for most metrics and a mix of the venues will be presented to examine the various metrics.

Two metrics that have significant correlations for the auditory-vocal task type are Mean Single-Glance Duration for the Road and Situational Awareness glance categories. These correlations are shown in Figure 6-25 and Figure 6-26. Here is can be seen that tasks with longer Total Activity Times tend to have longer single glances to the roadway and situational awareness locations. As Total Activity Time does not correlate to Task Duration well for auditory-vocal tasks this may represent a heightened workload state where subjects are staring as a result of cognitive tunneling due to task demands and then periodically taking longer glances to other locations.

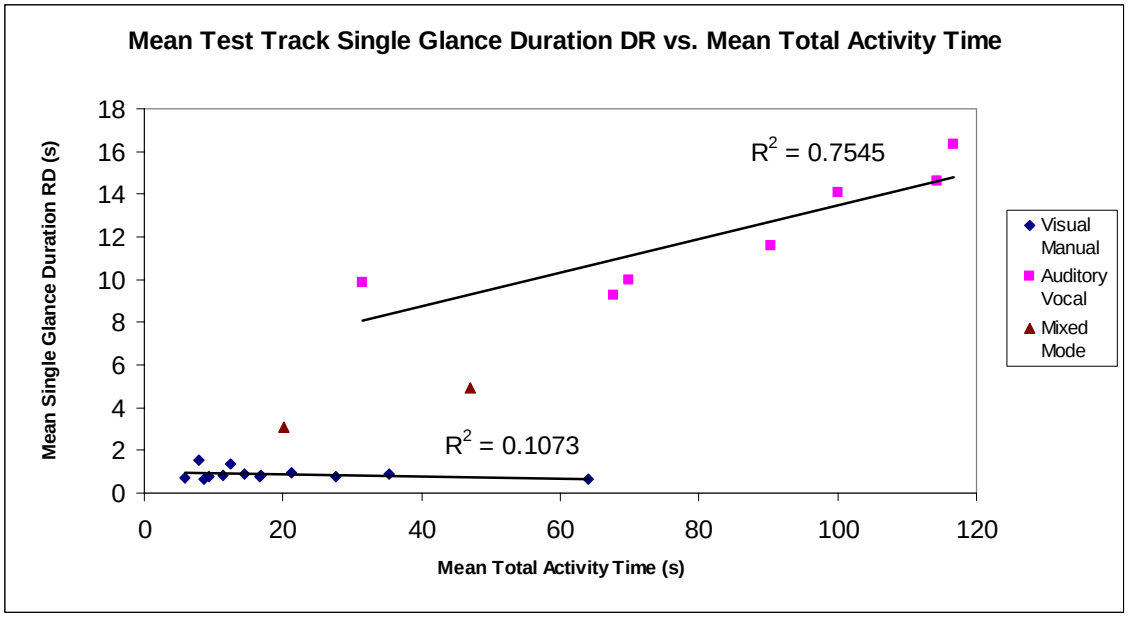


Figure 6-25. Test Track Mean Single-Glance Duration Road Versus Mean Total Activity Time

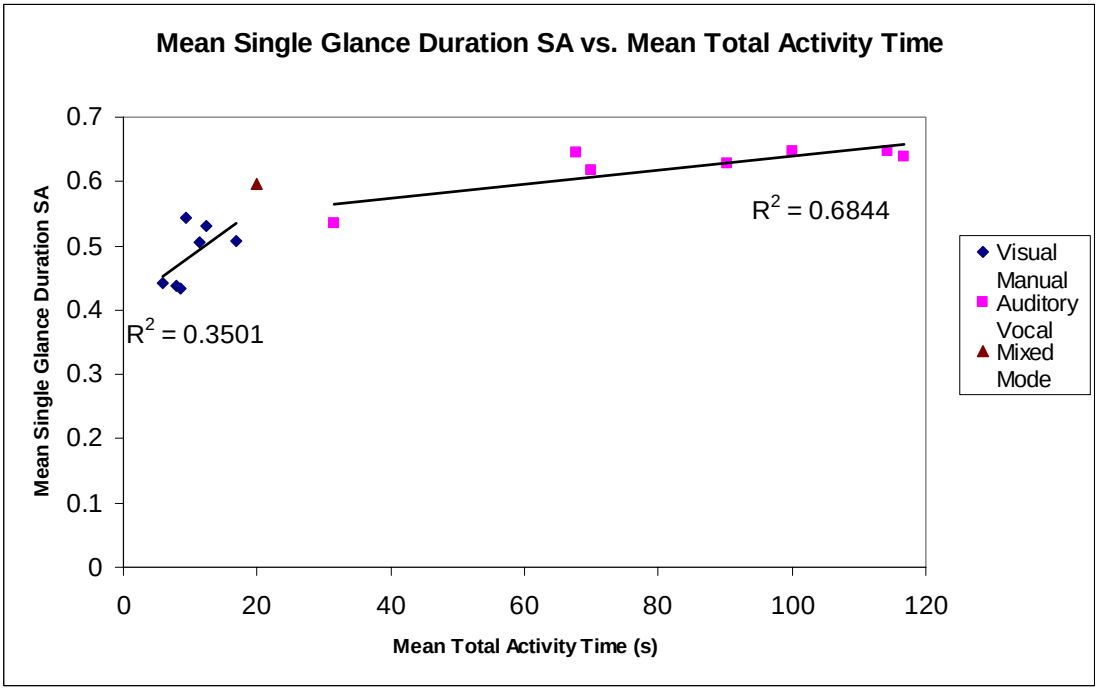


Figure 6-26. Mean Single-Glance Duration to Situational Awareness Versus Mean Total Activity Time

For the visual-manual task type a number of glance count metrics have significant correlations with Mean Total Activity Time. Figure 6-27 presents this correlation for the Test Track Mean Number of Glances to the Road glance category. While Destination Entry is an outlier in this graph, the correlation of the other tasks remains high. As expected from a time-related metric, tasks with longer Total Activity Times tend to have more glances to the road category.

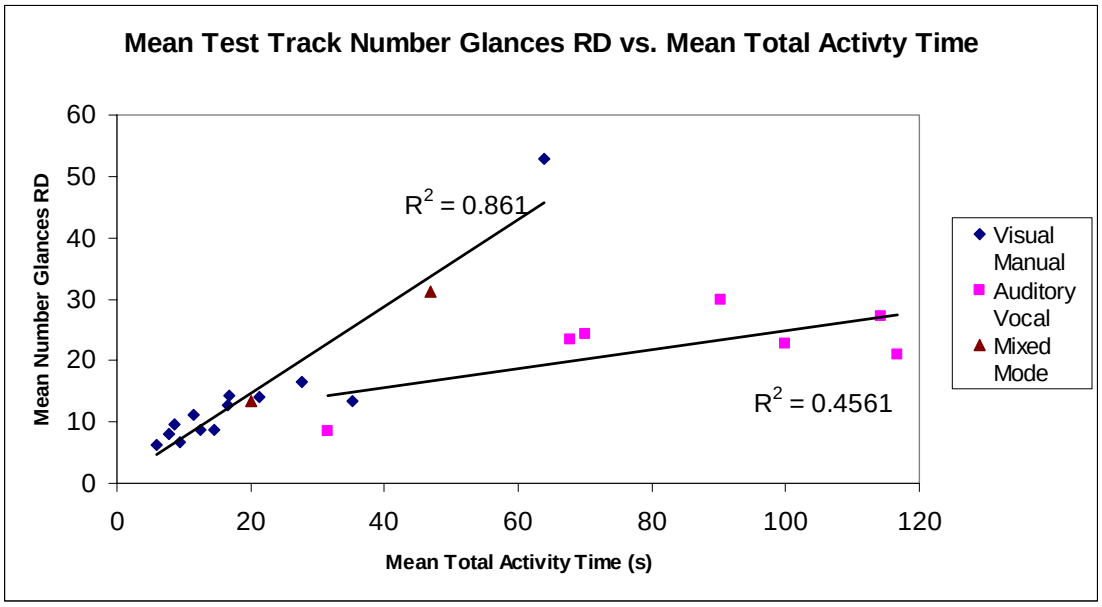


Figure 6-27. Test Track Mean Number of Glances to Road Versus Mean Total Activity Time

Figure 6-28 shows the correlation of Total Activity Time to Mean Number of Task Related Glances for the On-Road venue. As would be expected here the longer Total Activity Time tasks tend to have more task-related glances. The task with the most task-related glances here is Manual Dial, which Total Activity Time seems to be capturing by setting the task relatively far from the other shorter visual manual tasks. Logically, there is little relationship for the auditory-vocal tasks since these typically have few, if any, task-related glances.

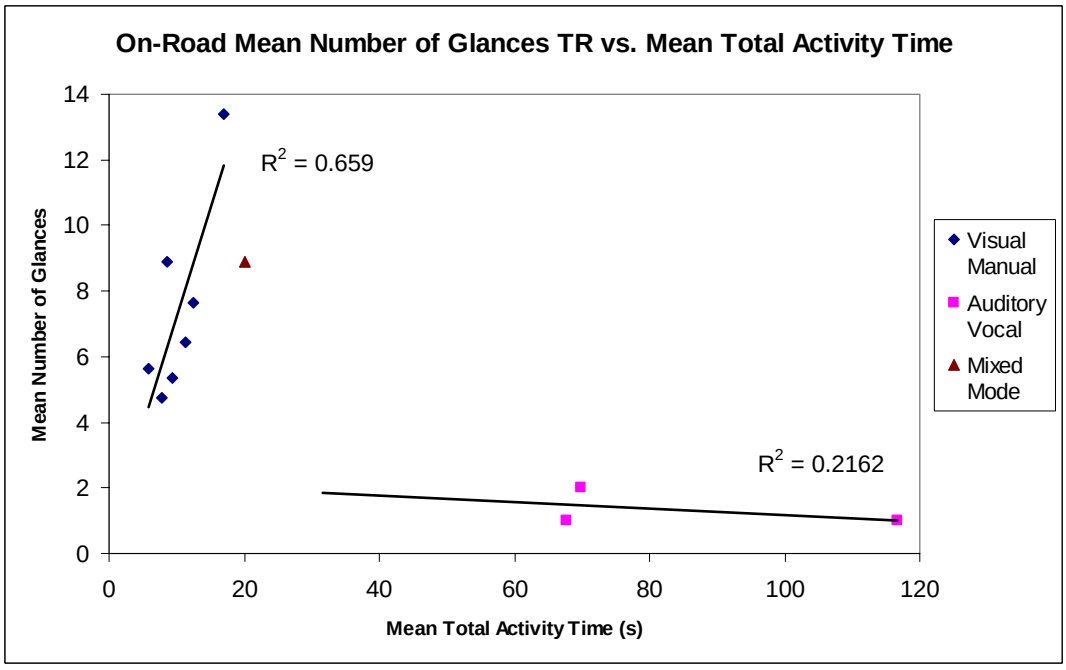


Figure 6-28. On-Road Mean Number of Glances Task Related Versus Mean Total Activity Time

Figure 6-29 presents the correlation of Total Activity Time with Mean Total Task Related Duration, which is the total duration of task-related glances during a task. As would be expected the correlation for auditory-vocal tasks is low and for visual-manual tasks it is high. With this graph, Destination Entry is again an outlier but within the cluster of visual-manual tasks there is good correlation between the metrics.

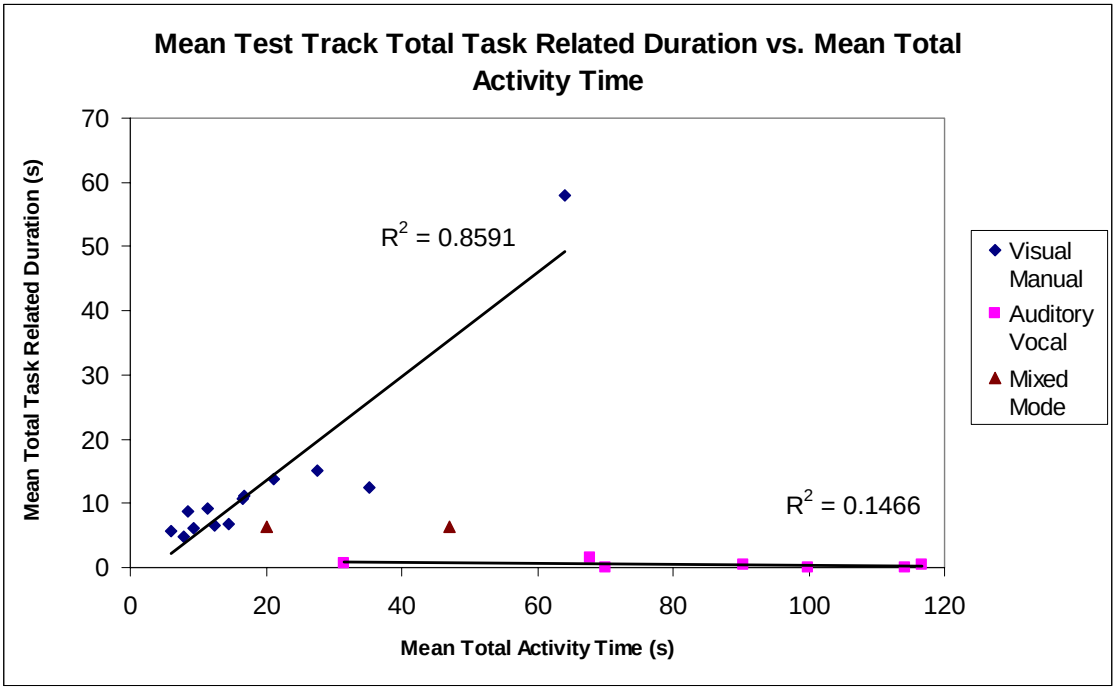


Figure 6-29. Mean Test Track Total Task Related Duration Versus Mean Total Activity Time

6.4.4 Task Activity Time Results

The Total Activity Time for a task model seems to be highly correlated with a number of other measures of secondary task performance ranging from task times to vehicle control and eyeglance metrics. Total Activity Time also exhibits the same short/long or visual-manual/auditory-vocal division of tasks as many other metrics. Correlations are typically higher for the entire task set due to this division however several metrics were still highly correlated for either the visual-manual or auditory-vocal task types. These relationships to other measures are far better than those of other metrics output by the modeling methodology.

Training, experience, and mediation of a final model would again help to improve the resultant models. Total Activity Time is still easily affected by task and modeling strategy differences. Detailed and accurate information on externally paced tasks is critical to assigning Activity Times. Numerous externally paced tasks were modeled and, due to differences in the inclusion of this pacing information and the variability in cognitive step construction, these tasks are always much less tightly clustered than the more observable self-paced tasks. These improvements may also produce more correlations between Activity Time and other metrics for the auditory-vocal task type. When examining these correlations it is often the case that one or two of the tasks are affecting the correlations (i.e. reducing them) below the level of significance. Corrections to the cause of these outlier tasks could enable more use of Activity Time as a surrogate for other metrics with auditory-vocal tasks.

Currently, Total Activity Time overestimates Static Task Time and Total Shutter Open Time while underestimating on-road and test track performance times. This would seem to indicate that the methodology as applied is indicative of something between single and dual task performance. The overestimation of the single task times may be due to including physical steps, such as Reach and Look caused by modeling from the point of view of the person driving. Since no effort was made to include any of the steps of driving in the models, it is not surprising that Total Activity Time underestimates real world performance times. A very interesting possibility is the development of either a driving model or dual task factor to represent the demands of driving. With such an addition to the methodology, a modeler could construct models from a Static Task mentality to more accurately represent just the task. Driving performance could then be predicted by the combination of the single task model and the driving component. This may be handled by the MRT component of the methodology but improvement to that component is needed first and will be discussed in the next section.

6.4.5 Dual Task Conflict Potential Results

The last metric yielded by the Task Steps Inventory methodology is DTCP, a measure based on Wickens' Modified MRT. The correlation chart in section 6.3 showed no significant correlations between DTCP and any other metrics. One way of rating tasks is to weight the DTCP by the time over which it is present. This is the idea behind multiplying DTPC by Total Activity Time. The sum of these products for each step is Activity Time*DTCP which showed numerous significant correlations in Section 6.3. When examined more closely, it can be seen that weighting DTCP in this way has lower correlation than Activity Time alone in all but a few cases. Rank orders when compared to longitudinal control metrics have higher correlations than the raw scores, similar to Total Activity Time results. This pattern does not hold with lateral control metrics however, rank order has lower correlation here than the raw scores. This leads one to hypothesize that DTCP is only adding noise to the Total Activity Time results and is thus not very predictive of task demand.

Figure 6-30 shows the relationship for Median TSOT and has an R^2 of 0.77 however as with other Occlusion comparisons, most all of this is due to the Destination Entry task. As with Activity Time the R^2 for this metric drops to 0.48 when Destination Entry is removed from the task set for Occlusion, thus time weighted DTCP is not highly predictive of Total Shutter Open Time.

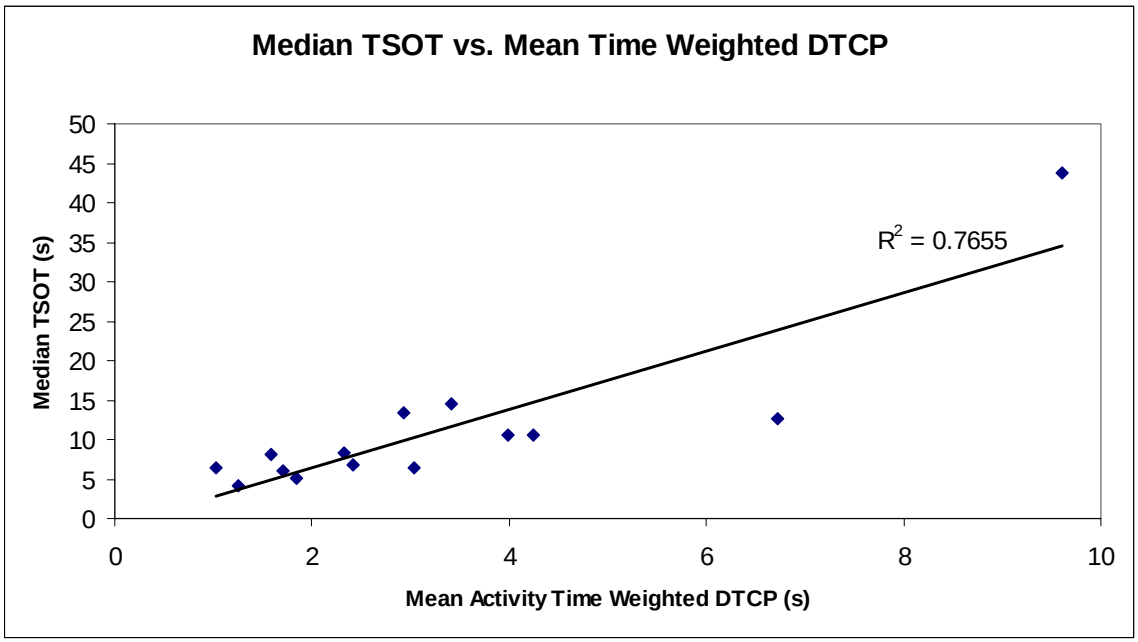


Figure 6-30. Median TSOT Versus Mean Activity Time Weighted DTCP

Given the similarity of DTCP results to those of Total Activity Time alone, and the lower correlations involved for DTCP, it is concluded that the current implementation of Modified MRT in the form of DTCP is flawed. Whether this is due to inconsistencies, as discussed in section 6.3, or to the calculation of DTCP itself is unknown at this time and is an area for future investigation.

6.5 Chapter References

Angell, L.S., Young, R. A., Hankey, J.M., and Dingus, T. A. (2002). An evaluation of alternative methods for assessing driver workload in the early development of in-vehicle information systems. SAE Proceedings, 2002-01-1981, Government-Industry Meeting, May 5, 2002, Washington, DC, USA

Card, S. K. (1983). User perceptual mechanisms in the search of computer menus. *Proceedings of Computer-Human Interaction Conference CHI, '82* (pp. 190-196). New York: Association for Computing Machinery.

Card, S., Moran, T. P., and Newell, K. (1983). *The psychology of human-computer interaction* (23-97). Mahwah, NJ: Lawrence Erlbaum.

Carroll, J.B. and Freedle, R.O. (1972). *Language Comprehension and the Acquisition of Knowledge*. Washington, DC: Winston.

Crowder, R. G. (1982). *The psychology of reading: An introduction* (p. 8-33). New York: Oxford University Press.

Deatherage, B. H. (1972). Auditory and other sensory forms of information presentation. In H. Van Cott and R. G. Kinkaide (Eds.), *Human engineering guide to equipment design* (pp. 123-160). Washington, DC: Government Printing Office.

Ericsson, K.A. and Simon, H.A. (1984) Protocol Analysis: Verbal Reports As Data. Cambridge, Mass: The MIT Press, From p 251.

Frost, G. (1972). Man-machine dynamics (pp. 227-310). In H. Van Cott and R. G. Kinkaide (Eds.), *Human engineering guide to equipment design* (pp. 311-344). Washington, DC: Government Printing Office.

Hankey, J. M., Dingus, T. A., Hanowski, R. J., Wierwille, W. W., and Andrews, C. (2000a, August). *In-vehicle information systems behavioral model and design support: Final report* (Report No. FHWA-RD-00-135). Washington, DC: U.S. Department of Transportation Federal Highway Administration.

Kintsch, W. (1974). *The Representation of Meaning in Memory*. Hillsdale, NJ: Erlbaum.

Seibel, R. (1972). Data entry devices and procedures. In H. Van Cott and R. G. Kinkaide (Eds.), *Human engineering guide to equipment design* (pp. 311-344). Washington, DC: Government Printing Office.

Wickens, C. D., and Hollands, J. (1999*). Engineering psychology and human performance* (3rd Edition). Upper Saddle River, NJ: Prentice Hall

7 Individual Differences and Driver Workload Metrics

7.1 Introduction to Individual Differences

A number of factors including subsidiary in-vehicle tasks, surrounding traffic, roadway environment, individual differences, and interactions between the factors are needed to explain an individual's driving performance. The effects of tasks and environment are discussed elsewhere in this report. This chapter focuses on an analysis of the differences among test participants, including age, gender, and performance on cognitive tests.

The first section of this chapter provides a brief overview of data processing and analysis. The next section covers the analysis of the effects of age on various laboratory measures and cognitive tests, and on-road and test track driving performance measures. This is followed by the effects of gender using the same format as the effects of age. The next section presents the results of two sets of cognitive tests and their ability to predict driving performance. Finally, the last section addresses the results of test participants' self assessments of tasks and abilities.

7.2 Description of Data and Processing

To prepare data sets for analysis, the on-road and test track data were processed and task epochs extracted. Next, the data were summarized by task occurrence, yielding statistics such as minimum, mean, median, maximum, standard deviation, and the amount of time a test participant held a particular value such as maximum speed or minimum range rate. A number of derived measures such as speed difference and speed change also were computed from the summarized data. The road and track data were then collapsed via simple means across all available replications yielding a test participant by task data set.

To reduce complexity of analysis and simplify understanding of results, the tasks were collapsed at the level of test participant by task, via simple means, into four task types: Just Drive, visual-manual, auditory-vocal, and mixed-mode. To reduce the number of analysis categories further, test participants were grouped into three age groups, young (20 to 39), middle (40 to 59) and old (60 to 79). For example, all 20 to 39 year old participants' individual Voice Dial and Delta Flight Information (Delta Flightline) tasks were collapsed into a data point for the mixed-mode task type and young age group. Due to a very skewed distribution and or little variation in individual differences cognitive test scores, these metrics were categorized prior to analysis as well. These categories were based on two or three levels, depending on the distribution, determined by below/above median value or approximately lower, middle, and upper third of the samples.

For the laboratory data, the raw data files were processed and task epochs extracted, then summarized by task occurrence. This yielded similar statistics to the vehicle data for STISIM and for miss rates, reaction times, percentage correct responses and durations for the other surrogate tests.

The laboratory data were then collapsed via simple means across replications for all measures yielding again a test participant by task data set. Data were then collapsed to four task types and an age category was created in the same way as vehicle data. In addition, age decade was also retained to allow for analysis of six categories of age as well. Individual differences cognitive test scores were categorized into two or three levels as was done with the vehicle data.

A number of analyses were done using these data. The first was a multivariate analysis of variance for repeated measures used to examine potential affects of age, gender, and task. Task type was treated as the within-subject effect; age and gender were between-subject effects; and vehicle metrics, one at a time, were treated as dependent variables. Second order interactions were examined as well. Post hoc tests were also employed to determine which particular classes were truly different whenever there was a significant main effect or interaction found.

Summary data sets and graphs with the means of significant effects of age and gender were then generated. These include age and gender by task and task type for both on-road and test-track data.

In a similar manner, multivariate analyses of variance were also conducted for individual differences cognitive tests. Here, task is the within-subject effect; one at a time the cognitive test scores were used as between test participants' effects; and the vehicle metrics were the dependent variables. Again, summary data sets and graphs were constructed, however, the classifiers are not age and gender but rather the categorized cognitive test score variables for road and track data.

A graphing convention was adopted that is used throughout this chapter. Results of an analysis of variance are presented for age and gender effects. In the graphs, a single, oversized data-point marker indicates that this point is statistically significantly different from the other data point or points for that task or task type. If there are two such markers, this indicates the same for just those two means. Task types are assigned in this chapter as follows, type 1 is the Just Drive task, type 2 is the collapsed visual-manual task set, type 3 is the collapsed auditory-vocal task type and type 4 is the mixed-mode task type.

7.3 Age Effects

With the growing aging population in the United States., age and its effect on driving performance are increasingly more important. Numerous studies have been conducted to examine the effects of age, both young and old, on various aspects of driving. With current advancements in vehicle telematics and devices, it is critical to understand how these devices will affect drivers' performance. This need is even more important with older drivers who may be more sensitive to additional attentional, cognitive, and complex motor skills demands. This section details some differences in performance of three age groups on surrogate tests, on-road and test-track driving performance.

7.3.1 Laboratory Age Effects

Divergence of the R-Metric for task type 4 in Figure 7-1 comes from older test participants having higher TSOTs for all tasks and the youngest test participants having the lowest task 2 time. The youngest test participants also had the highest task 4 time with TSOTs slightly less than the trend with other age groups.

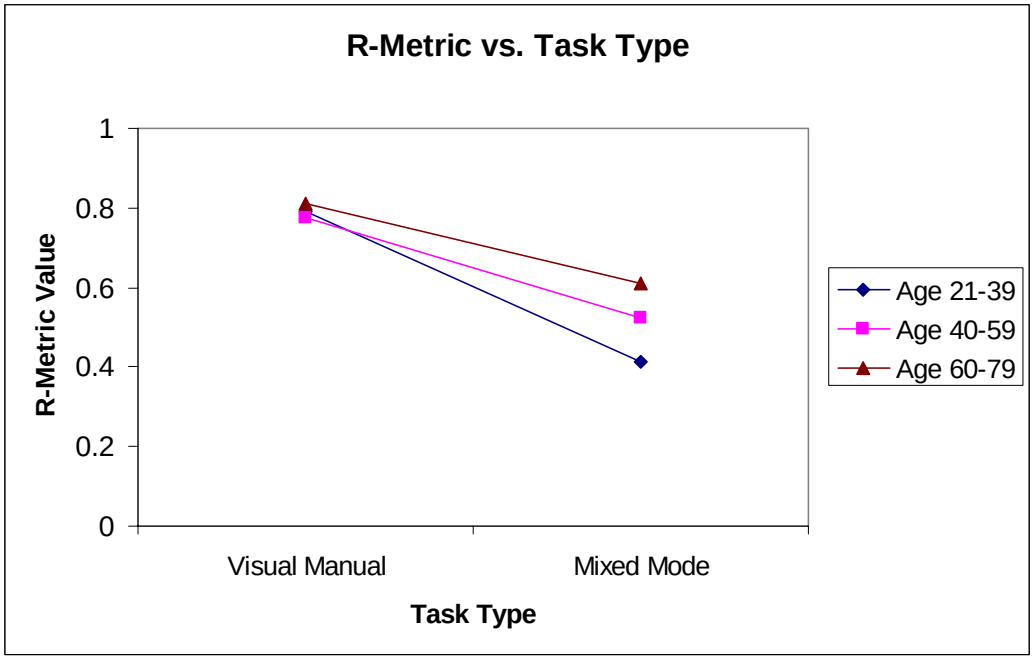


Figure 7-1. Age Effects – Laboratory Test Participants R-Metric by Task Type

STISIM mean task duration by age group is presented in Figure 7-2. This graph shows that mean task times for the three age groups range from 26 to 40 seconds for visual-manual tasks but are more consistent for the mixed-mode type with a range of 76 to 82 seconds.

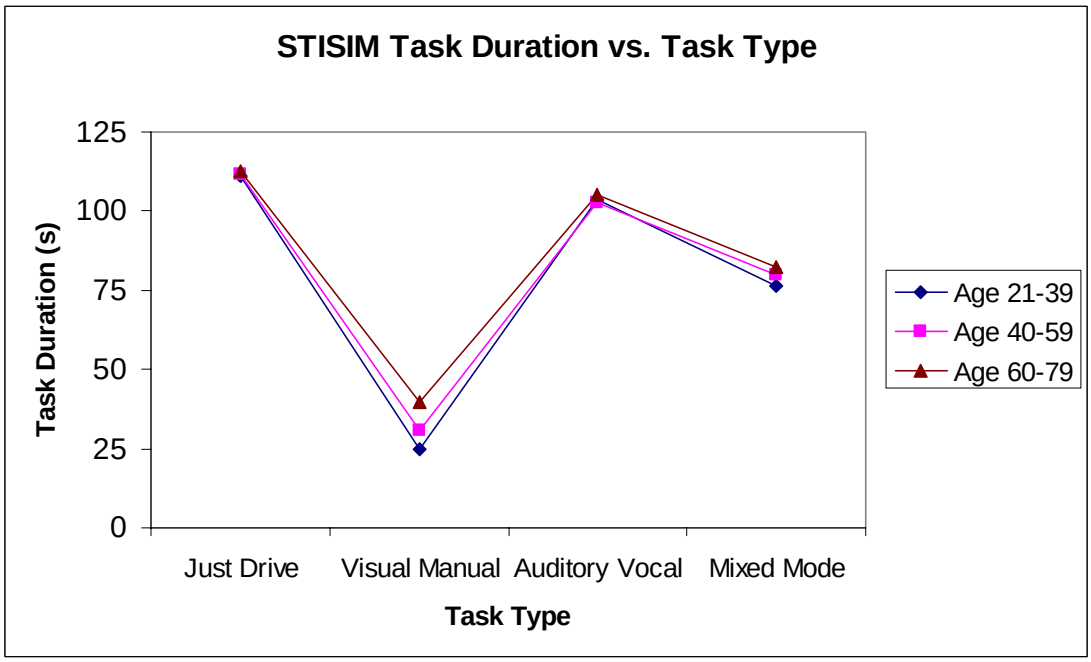


Figure 7-2. Age Effects – STISIM Task Duration by Task Type

Lateral control shows a significant age effect for mean standard deviation of lane position (SDLP), as well as for lane exceedance counts. A representative graph for STISIM mean time out of lane is shown in Figure 7-3. While the oldest age group has the highest SDLP for all task types, the difference from other groups is only significant for the visual-manual task type. When lane exceedance data is examined, however, a different trend emerges. When a lane exceedance occurs during Just Drive or auditory-vocal task types, the oldest test participants spend the least time outside their lane. For the two tasks with a manual component however, the oldest test participants spend the most time outside their lane. This same trend is also observed for lane exceedance touch and cross counts.

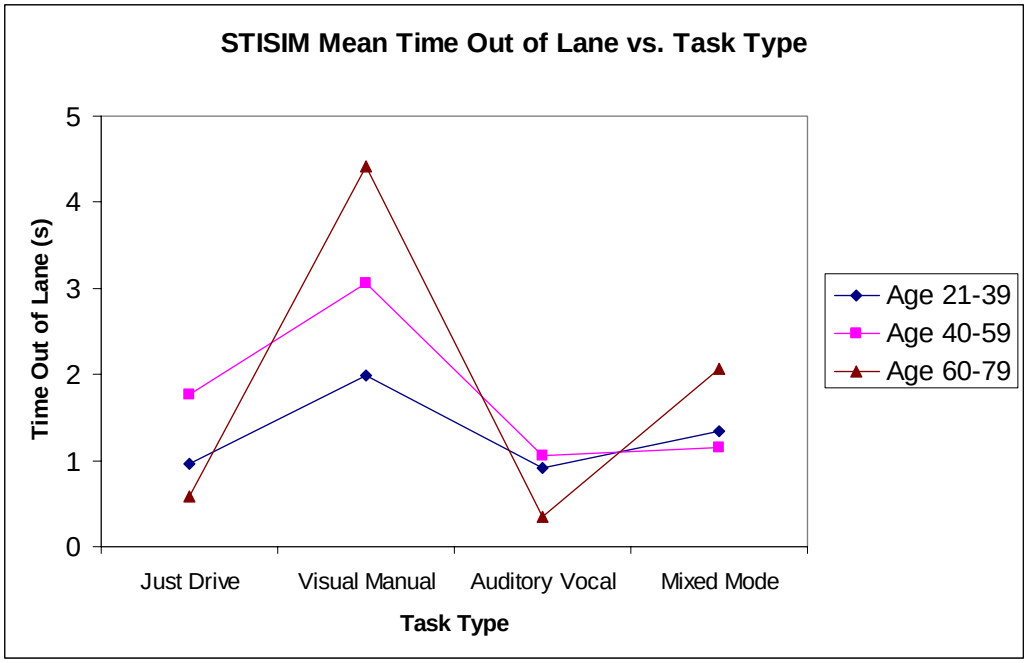


Figure 7-3. Age Effects – STISIM Time Out of Lane by Task Type

Figure 7-4 shows the age effect for STISIM mean speed difference. Older test participants had a larger range of vehicle speed for all tasks, but the trend was most pronounced for task types 2 and 4, which had a manual component. It is important to note that these results were also seen in on-road and test-track performances, but the effect was larger for all ages, possibly due to the lack of vestibular cues, roadway conditions or "platooning effects" with the driving simulator.

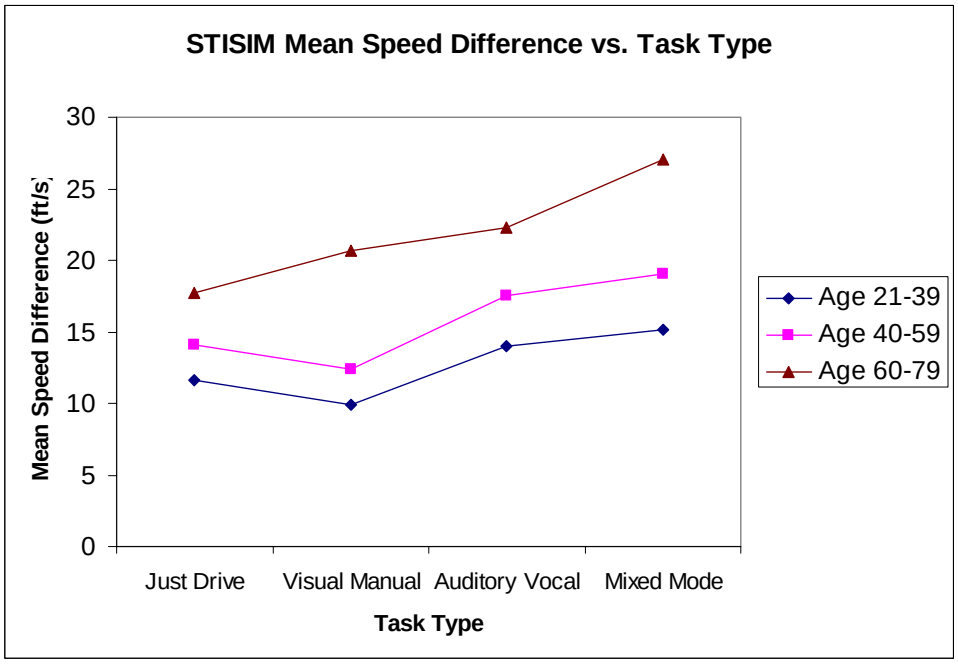


Figure 7-4. Age Effects – STISIM Speed Difference by Task Type

The age effect for PDT-Alone miss rate, seen in Figure 7-5, is quite similar to the one shown previously for lane exceedance data. The oldest test participants have miss rates comparable to the other age groups for auditory-vocal and Just Drive tasks, but these rates are significantly higher than other age groups for visual-manual and mixed-mode task types.

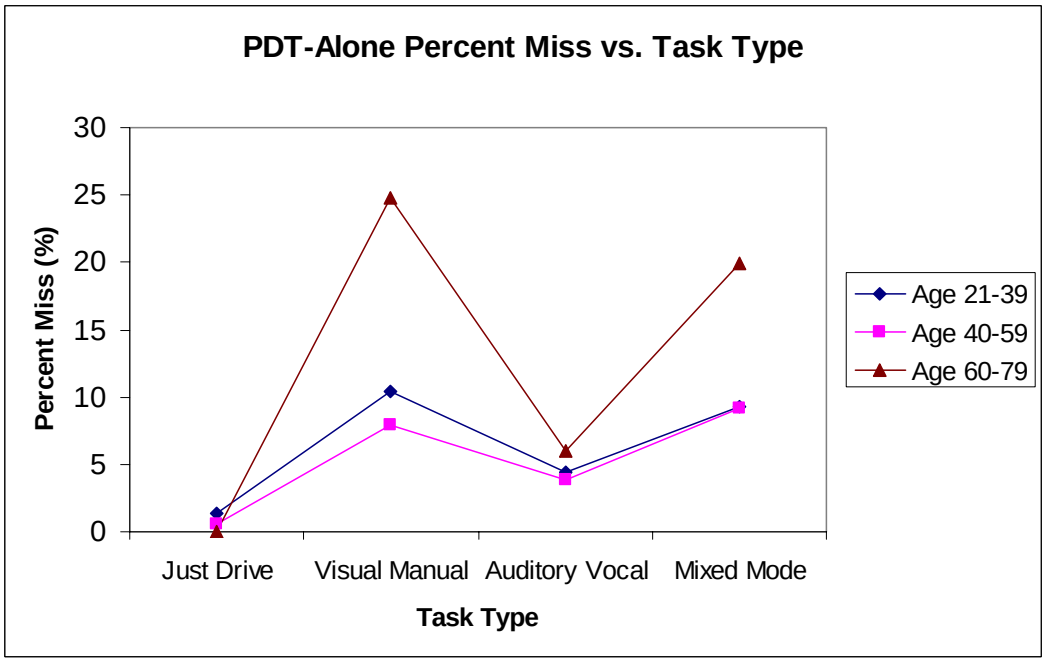


Figure 7-5. Age Effects – PDT-Alone Miss Rate by Task Type

Figure 7-6 shows the PDT miss rate for PDT with STISIM. There are now higher miss rates and more variability between age groups for all tasks, however, the age effect is still present and most significant for task types with a manual component. Interestingly, there does not exist a corollary effect with reaction time for either of the PDT surrogates.

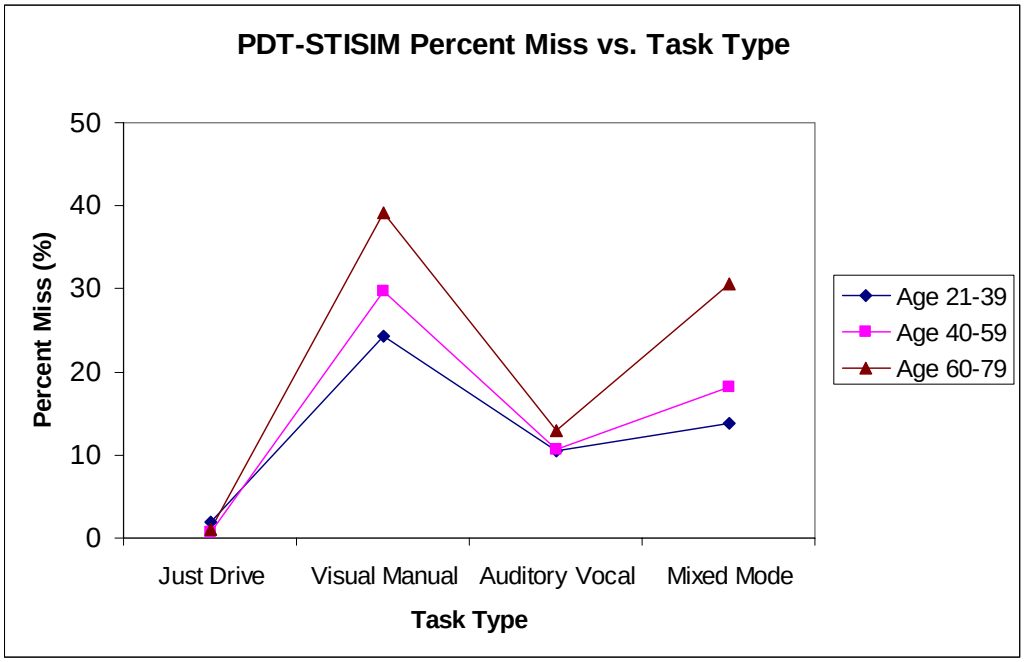


Figure 7-6. Age Effects – PDT with STISIM Miss Rate by Task Type

The Sternberg memory task also presents visual stimuli so one would expect to see similar age effects to those seen with PDT. Figure 7-7 shows the age effect for percent missed detections in the Sternberg test. Here, the same manual component tasks have significantly higher rates of missed detections. Sternberg error rates also show the same trend both between tasks and between age groups.

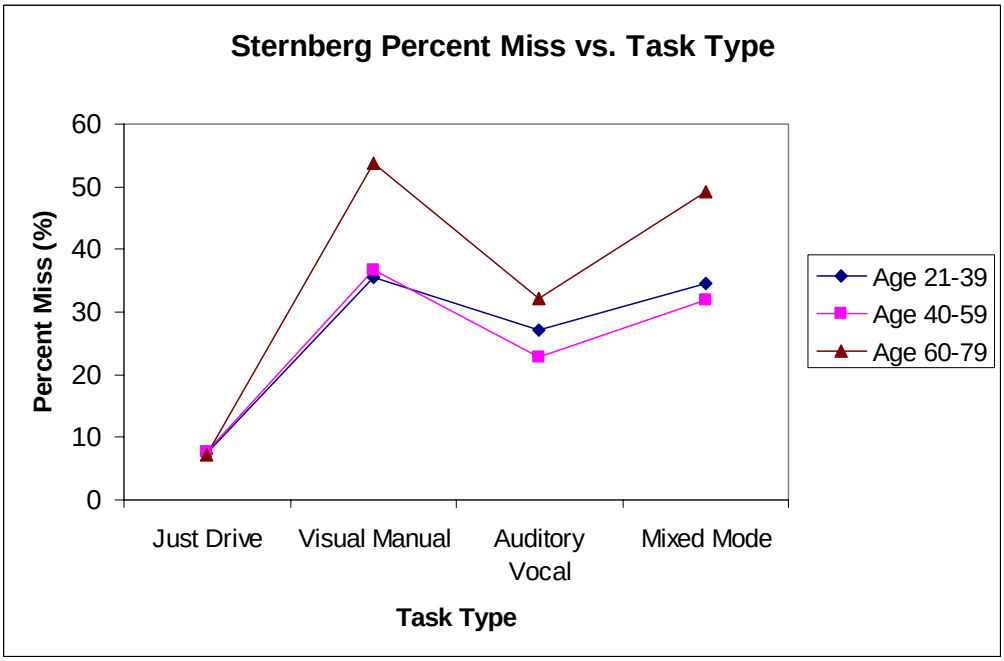


Figure 7-7. Age Effects – Sternberg Percent Miss by Task Type

In Table 7-1 the mean scores for a number of cognitive test metrics are presented by age group. For the Patsys Manikin and Grammatical Reasoning tests, scores follow age with the youngest test participants having the best performance. For Useful Field of View measures, the middle age group scores better than the youngest. The Baddley Dual Task score indicates that on average, the youngest test participants perform better when dual tasking than when performing a single task. This may offer some explanation for younger test participants having poorer driving performance for the Just Drive task than the middle age group for a number of measures seen elsewhere. In all cases, however, the oldest age group has the lesser performance on this set of cognitive tests.

Table 7-1. Cognitive Tests Scores by Age Group

Cognitive Test Metric	Scoring	Age Group		
		1 (21-39)	2 (40-59)	3 (60-79)
Patsys Manikin Percent Correct	Higher = better	95.333	94.938	80.067
Patsys Manikin Mean Response Time	Lower = better	1.942	2.595	2.667
Patsys Grammatical Reasoning Average Percent Correct	Higher = better	78.867	77.688	56.933
Useful Field of View Processing Speed	Lower = better	26.250	18.042	24.756
Useful Field of View Divided Attention	Lower = better	40.459	35.292	113.422
Useful Field of View Selective Attention	Lower = better	87.292	86.178	201.689
Baddeley Dual Tasking	Higher = better	102.490	98.009	94.121

To summarize the age effects for laboratory data:

- For lateral control in STISIM, the oldest test participants show slightly higher SDLPs with the most pronounced effect for the visual-manual task type. This carries over to lane exceedance counts and time out of lane with a larger difference between age groups and between the oldest group and the other two. The magnitude of these measures does not follow task duration as is seen in some other cases.
- The oldest test participants differ significantly from the other two groups with regard to STISIM speed difference. They also exhibit a difference between task trend for tasks with a manual component. For task types 2 and 4, the oldest test participants have more variation, whereas the other two groups have less for these than for the other task types.
- The oldest test participants have the highest PDT miss rates of the age groups for visual-manual component task types while being nearly equal to the other groups for Just Drive and auditory-vocal task types. Reaction times for PDT do not exhibit significant differences among the age groups. These hold for both PDT Alone and PDT with STISIM.
- The oldest test participants have the highest percent missed events and error rates for Sternberg memory task for all but the Just Drive task type.
- Age effects for the cognitive test scores vary, but generally, the oldest test participants have the poorest performance on these tests. The middle-age group can be nearly equally spaced or closer to either the oldest or youngest test participants depending on the cognitive measure being examined.

7.3.2 On-Road Age Effects

Figure 7-8 shows the Standard Deviation of Lane Position for the four task types in type series based on age groups. An analysis of variance shows that there is a significant task effect on SDLP. This analysis also shows a statistically significant degradation in lane keeping and an increase in variability of lane position for the 60 and 70 year old age group for the visual-manual task type.

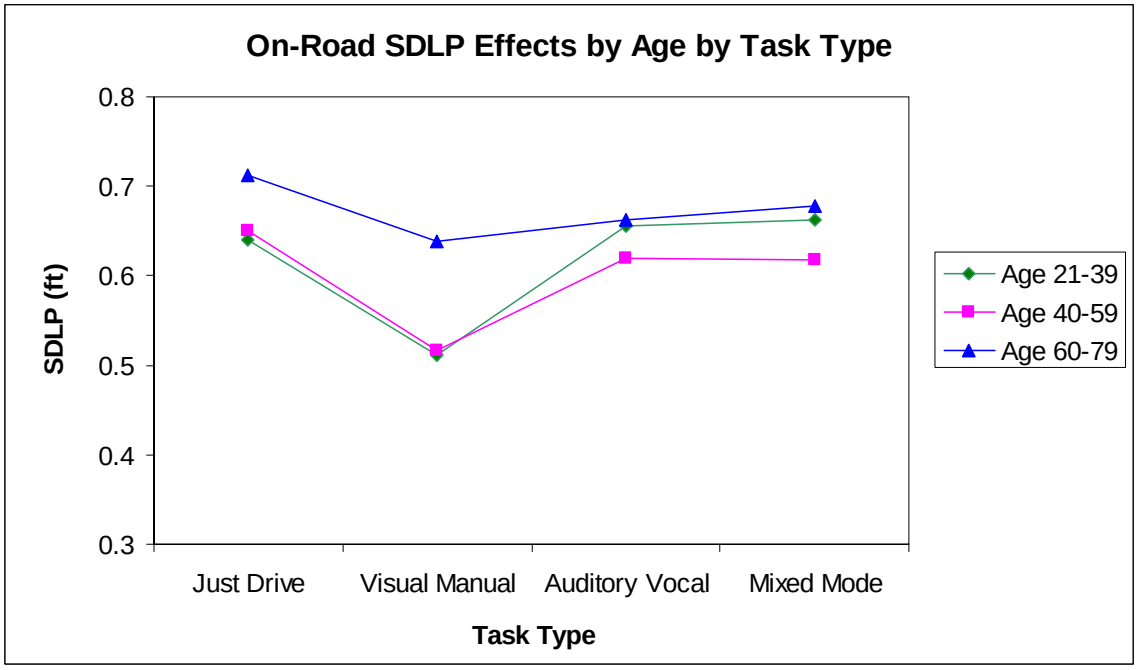


Figure 7-8. Age Effect – On-Road Standard Deviation of Lane Position

Figure 7-9 shows the mean summary statistics of range, by task type, for each age group. The minimum, mean, and maximum range statistics all show significant task effects. While there are significant differences between the age groups for all of the statistics, the general trend between tasks is unchanged across age groups. Minimum range shows the only significant difference with the auditory-vocal task type, indicating that older test participants are not following as closely during these tasks. Mean range indicates the same behavior for older drivers that minimum range shows. Maximum range, however, shows less variability between tasks with older test participants than with other age groups. Maximum range also shows that older test participants are consistently falling farther back from the lead vehicle than the other age groups.

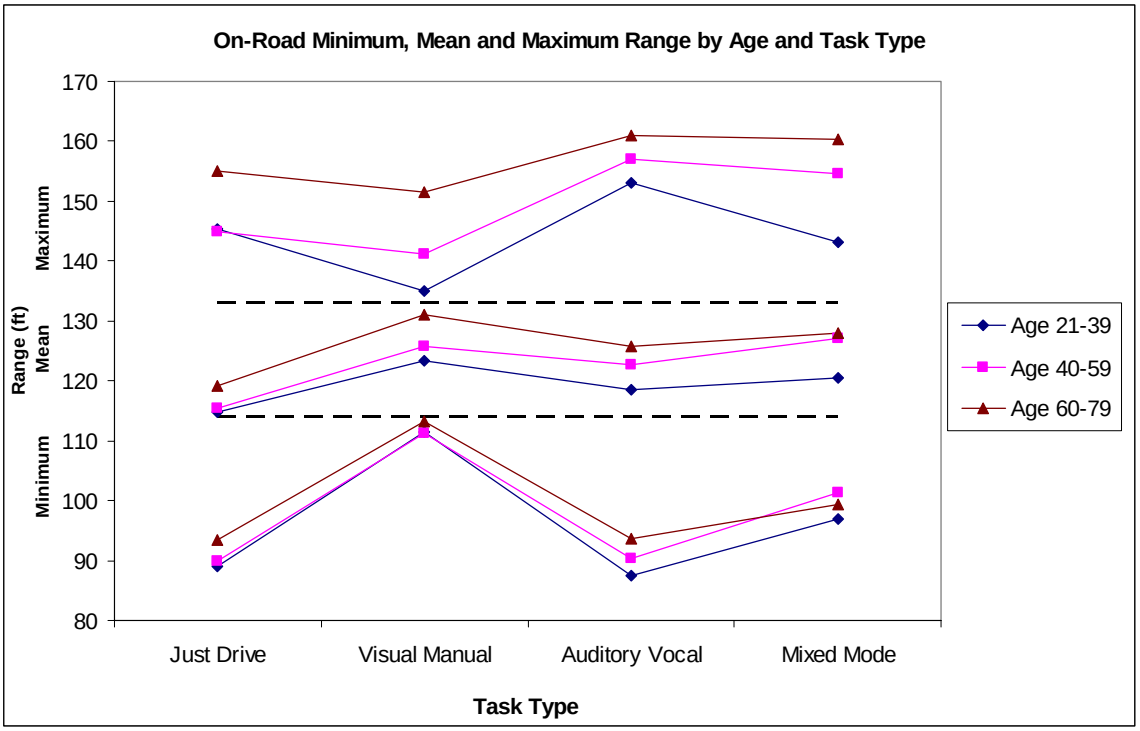


Figure 7-9. Age Effect – On-Road Range

Figure 7-10 shows the effect of age group on the standard deviation of range for the four task types. An analysis of variance showed significant task, age, and gender effects as well as task-by-age and age-by-gender interactions. The Just Drive and auditory-vocal tasks are closely grouped between age groups and are very similar in magnitude to each other. An interesting difference is seen in the task types with a manual component showing the same spread of the age groups, with older test participants having the most variability in range.

Figure 7-11 presents the minimum, mean, and maximum values of range rate by task type and age group. For all three measures, the visual-manual and mixed-mode tasks show significant differences. This indicates less stability in vehicle following with the older test participants having higher closing and receding rates and younger test participants having lower rates.

Figure 7-12 indicates that there is more variability in range rate for visual-manual and mixed-mode tasks for older test participants. The 20 and 30 year old group shows the least variability with visual-manual and mixed-mode tasks as well.

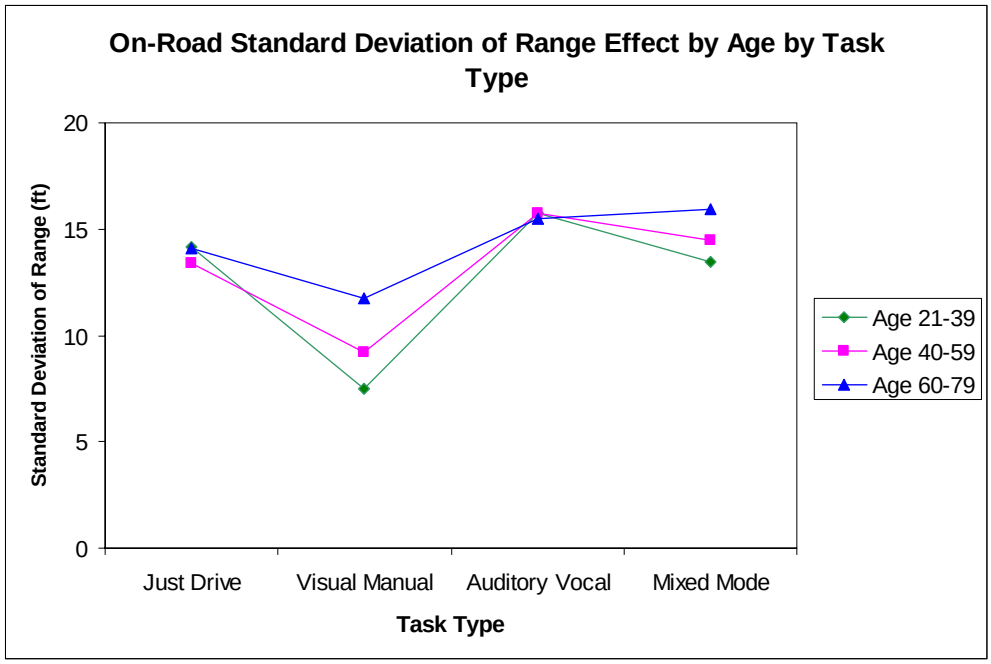


Figure 7-10. Age Effect – On-Road Standard Deviation of Range

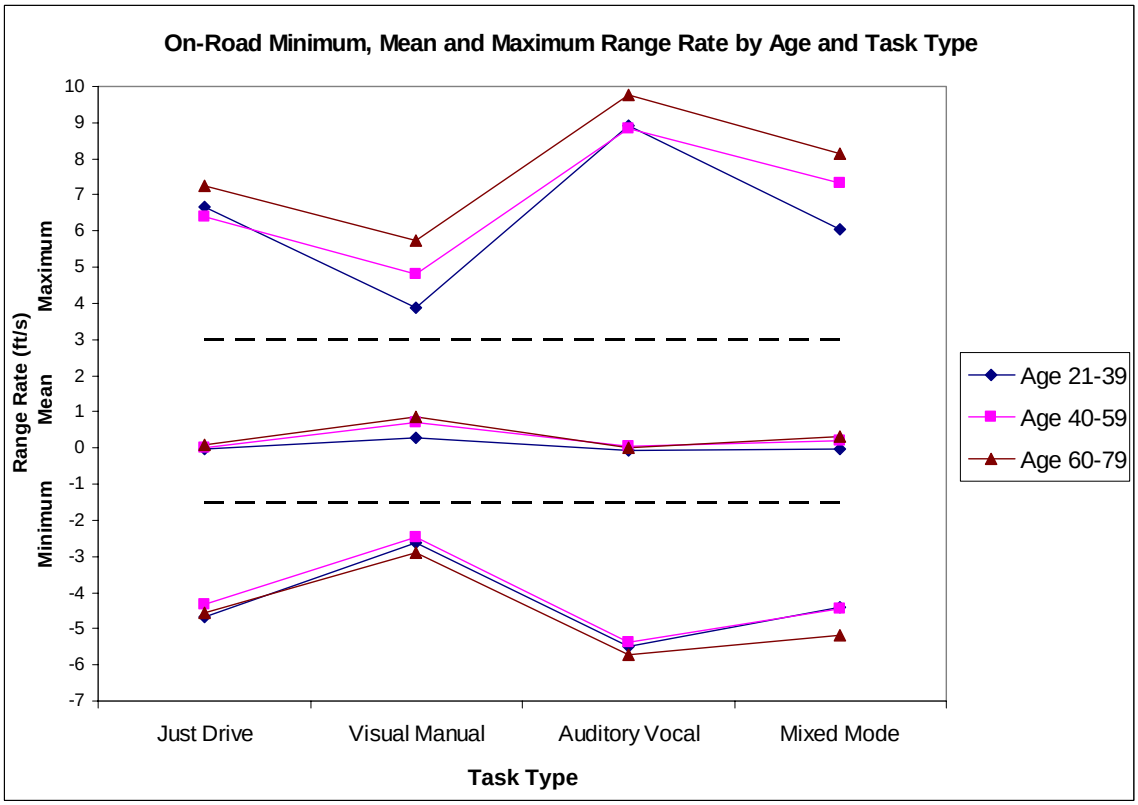


Figure 7-11. Age Effect – On-Road Range Rate

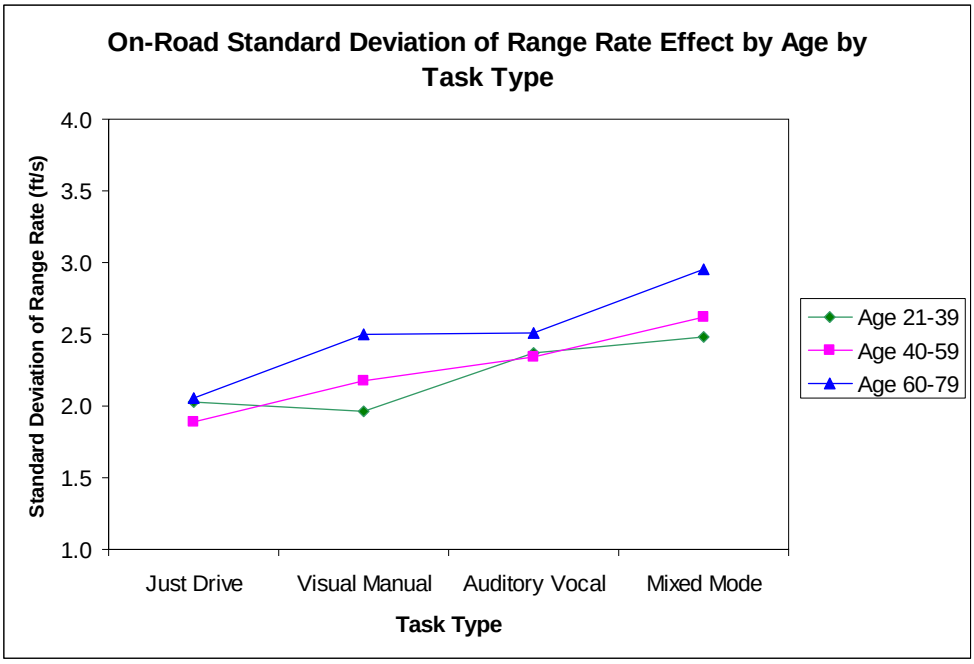


Figure 7-12. Age Effect – On-Road Standard Deviation of Range Rate

Figure 7-13 shows the mean summary statistics for speed by age and task type. Previously, maximum range and range rate showed the most differences between task types and age groups. In this graph, differences are seen in minimum speed, indicating that younger test participants slow less than older test participants. Generally, test participants are slowing less during the shorter visual-manual tasks than for other task types, and this trend is consistent across age groups and has an inverse trend in maximum speed.

Figure 7-14 shows the age effect on standard deviation of speed. As with other standard deviations for longitudinal measures, older test participants have more and significantly different variability on the visual-manual task type. It is interesting to note a trend across tasks for this measure, which is most pronounced in the older test participants. The variability over time of speed is greater for the mixed-mode task type, which as a group, is shorter in duration than the auditory-vocal task type.

Figure 7-15 shows not only significant task effects but also age effects for speed difference, defined as maximum minus minimum speed. In this graph it can be seen that older test participants have more variability in speed with significant differences from the other age groups for the task types with a manual component. The between task trend in speed difference can be seen to grow with task duration and is very consistent across age groups.

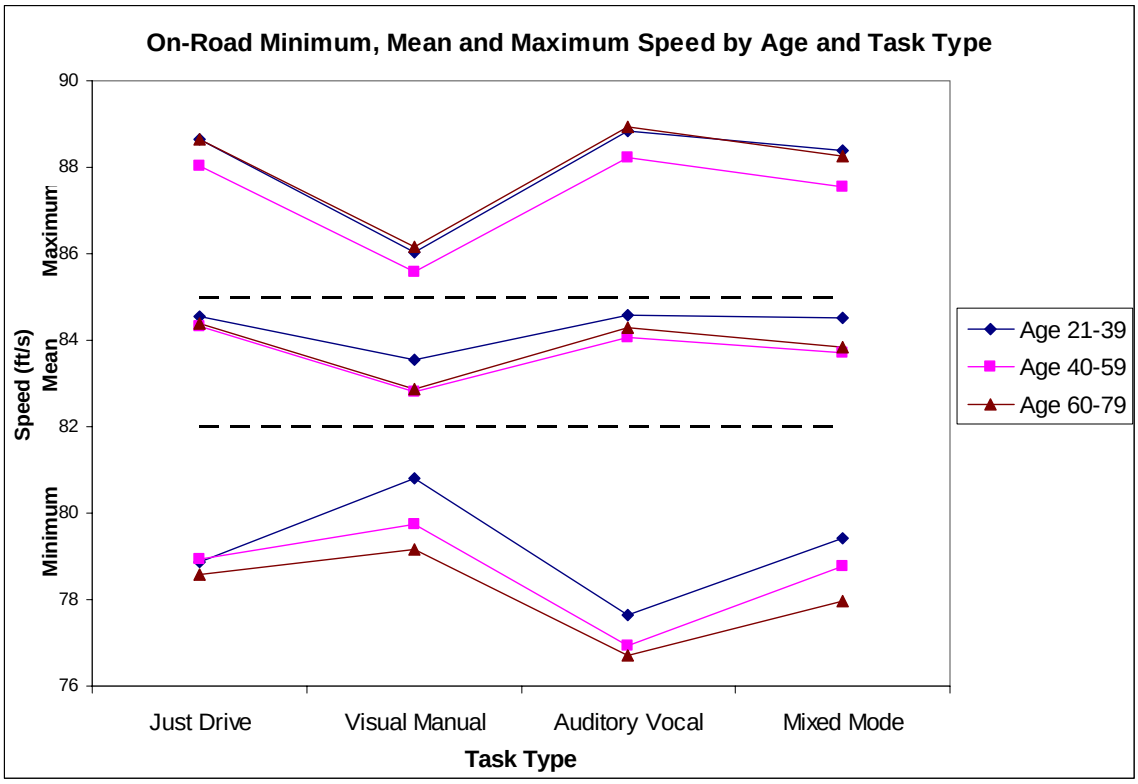


Figure 7-13. Age Effect – On-Road Speed

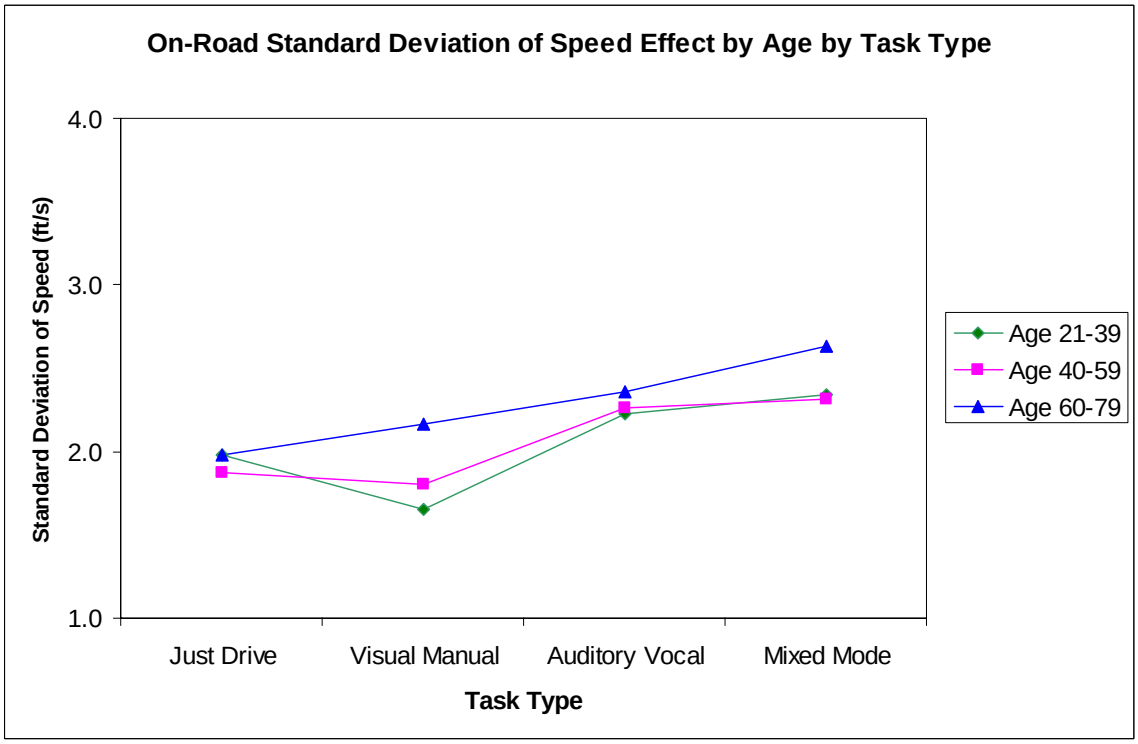


Figure 7-14. Age Effect – On-Road Standard Deviation of Speed

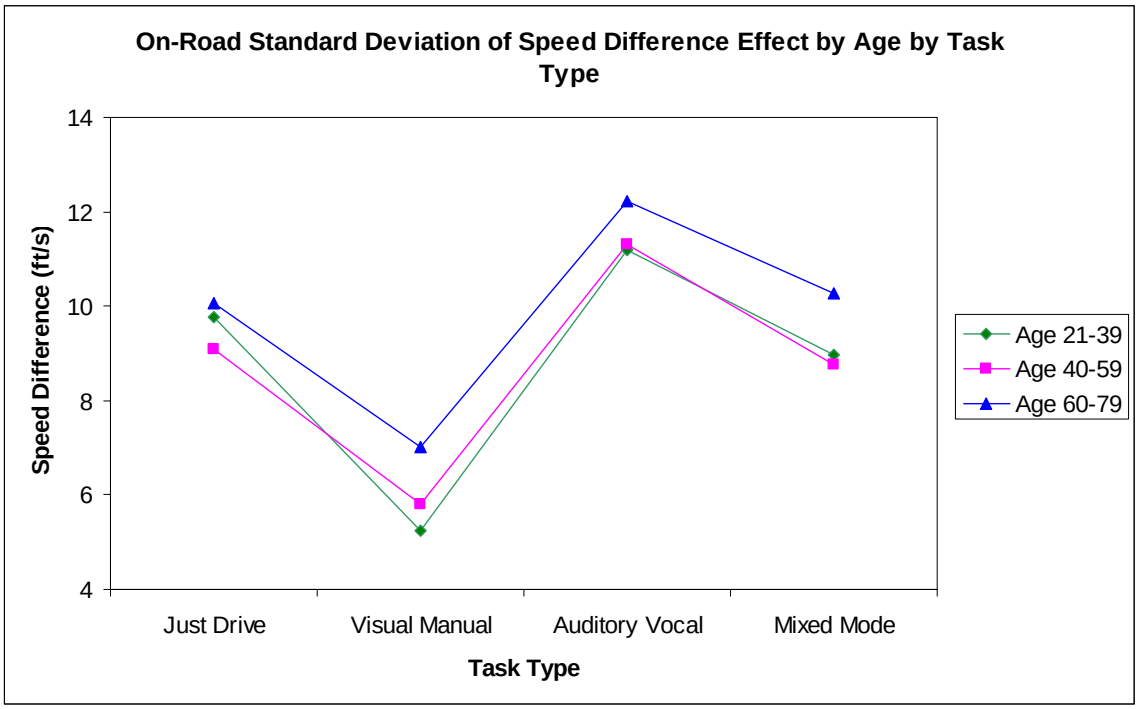


Figure 7-15. Age Effect – On-Road Speed Difference

The SDLP shows little variation in general between task types for the younger age groups. The oldest test participants though, have the highest values of all three groups for both the Just Drive and visual-manual task types and values nearly equal to the other groups for the remaining tasks.

The oldest test participants tend to have the most variability in longitudinal control and their difference from the other groups tends to be greatest for the task types with a manual component. This includes lower minimum speeds and higher maximum range and range rates.

The oldest test participants have higher standard deviations of range, range rate, and speed than the others, primarily for the manual component task types. While range exhibits a trend that follows task duration for this statistic, the other two break this trend with the mixed-mode task having a standard deviation for all age groups that is at or above that of the much longer auditory-vocal task type.

Typically, the magnitude of any of the longitudinal measures varies in direct relation to task duration for all task types except Just Drive. The exception to this trend is Just Drive, which typically is the lowest and the mixed-mode task type, which is often the highest.

While the differences between age groups are often statistically significant, the absolute difference is often rather small. Determination of practical significance as well as future potential analysis could be the subject of future research.

7.3.3 Test Track Age Effects

Figure 7-16 presents the age effect, for the test track venue on SDLP. In addition to significant task and age effects, there is a significant task by age interaction. Though practically the differences are small, a significant difference exists for older test participants with all but the auditory-vocal tasks. Visual-manual and mixed-mode task types included more complex tasks for the test track than for the road. Given the new tasks, an accentuation of the manual component effects seen previously for older test participants is not unexpected. The significant difference for Just Drive, however, is unexpected for the test track venue. The interaction of task and age is also interesting since the youngest age group is shifting in relative position to have the better performance on the task types with a manual component.

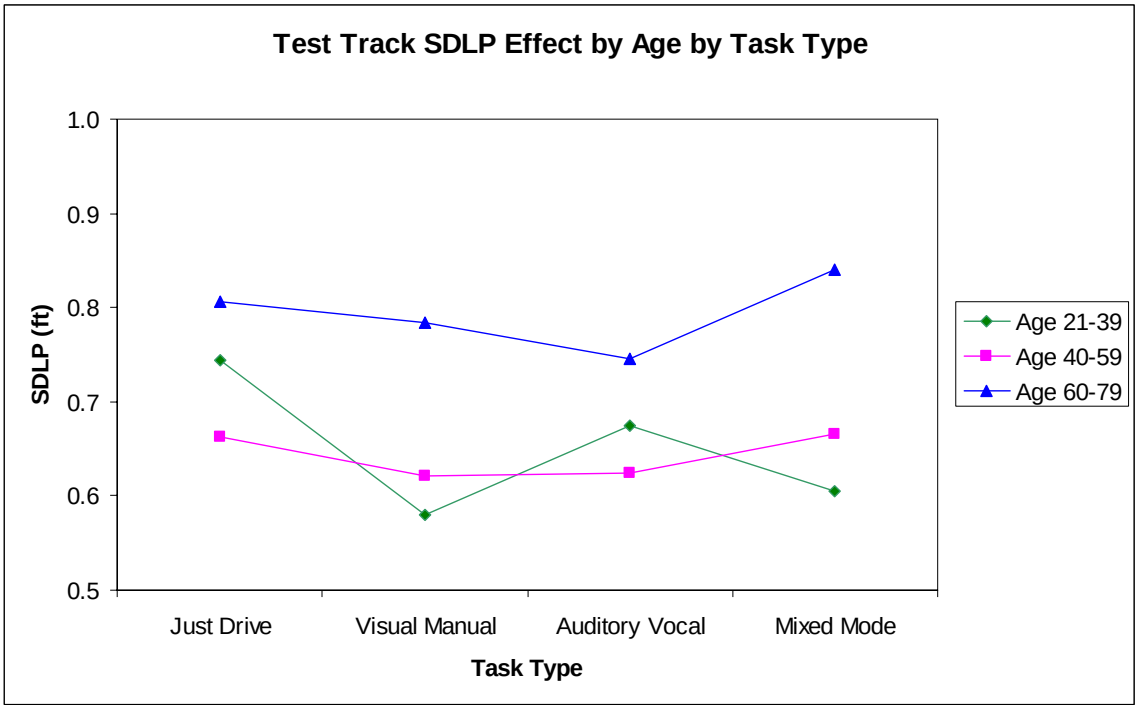


Figure 7-16. Age Effect – Test Track Standard Deviation of Lane Position

Figure 7-17 shows the values of summary statistics for range by age group and task type. While there is a significant task effect for all these measures, the interesting statistic is maximum range. In this venue, maximum range shows two significant interactions, task by age and age by gender. The three age groups are similar in minimum range and generally follow an increasing trend according to task length. Mean range shows less consistency among the age groups and while it does not possess a significant age effect, the mean range is different between young and old age groups for the auditory-vocal tasks. The significant age effects and interactions come with maximum range. There are different patterns between tasks for each of the age groups. Typically younger test participants fall back less from the lead vehicle while older test participants fall back more. Within the older test participants, maximum range is increasing from the shortest visual-manual task type to the auditory-vocal and mixed-mode task types. Given that the mixed-mode task type has a shorter duration than the auditory-vocal task type, this trend may not be completely dependent on task duration for older test participants.

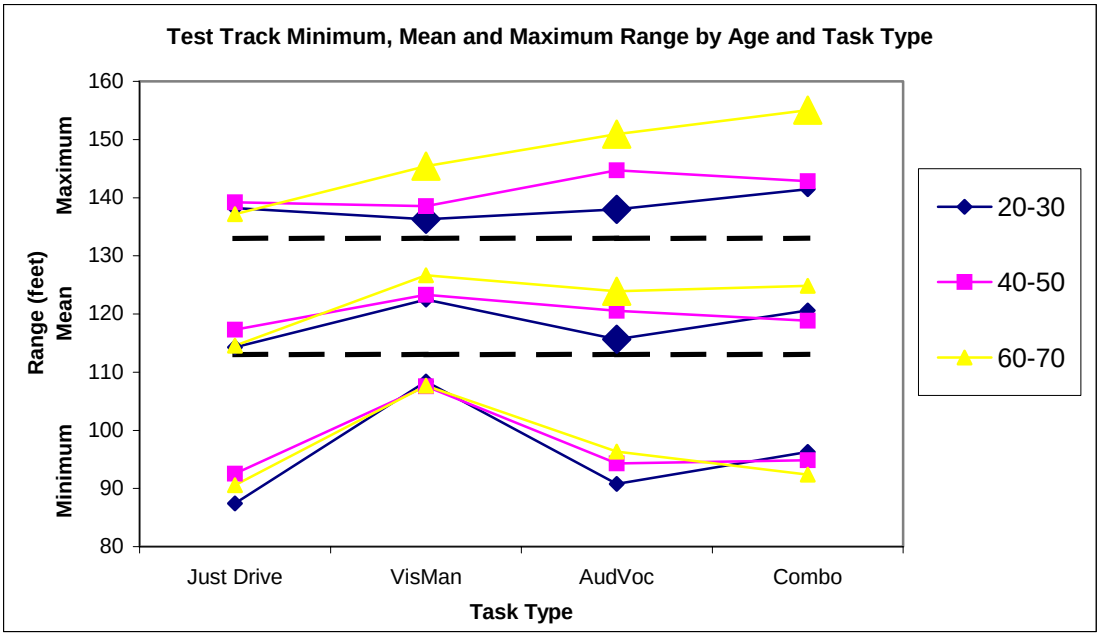


Figure 7-17. Age Effect – Test Track Range

Figure 7-18 shows the age and task effects on standard deviation of range. This graph shows the age by task interaction again, with younger test participants having more variability in range for the Just Drive task than for other task types. Older test participants also have significantly different measures for visual-manual and auditory-vocal tasks. The position of older test participants for the mixed-mode task also indicates that the addition of the Delta Flight Information (Flightline) task significantly affects the variability of range for this age group.

Figure 7-19 shows the minimum, mean, and maximum values of range rate by task for each of the age groups. Mean range rate is not significantly affected by age in this venue. Minimum range rate, however, is affected by age with a difference for the older test participants, again with the mixed-mode task. A lower value of range rate here indicates that the older test participants are surging forward at a greater rate to catch up to the lead vehicle than are the other age groups. Maximum range rate also shows significant differences by age groups. Younger test participants have significantly lower maximum range rates than older test participants. This means that the older test participants are falling back from the lead vehicle at higher rates than the youngest age group. The trend of greater variability within older test participants and the mixed-mode task is a break from the trend in the on-road venue where Voice Dial was the only mixed-mode task type. Since auditory-vocal and Just Drive task types both have much longer task durations than the mixed-mode task type, this tends to indicate that for this age group/task mixed-mode, variation is not driven strictly by task duration.

Figure 7-20 shows the task and age effects for standard deviation of range rate. As with other measures, the auditory-vocal task type is similar to Just Drive for most test participants. The task types with a manual component, however, show much more variation within the older test participants, again with the mixed-mode task having the most variation in this longitudinal measure.

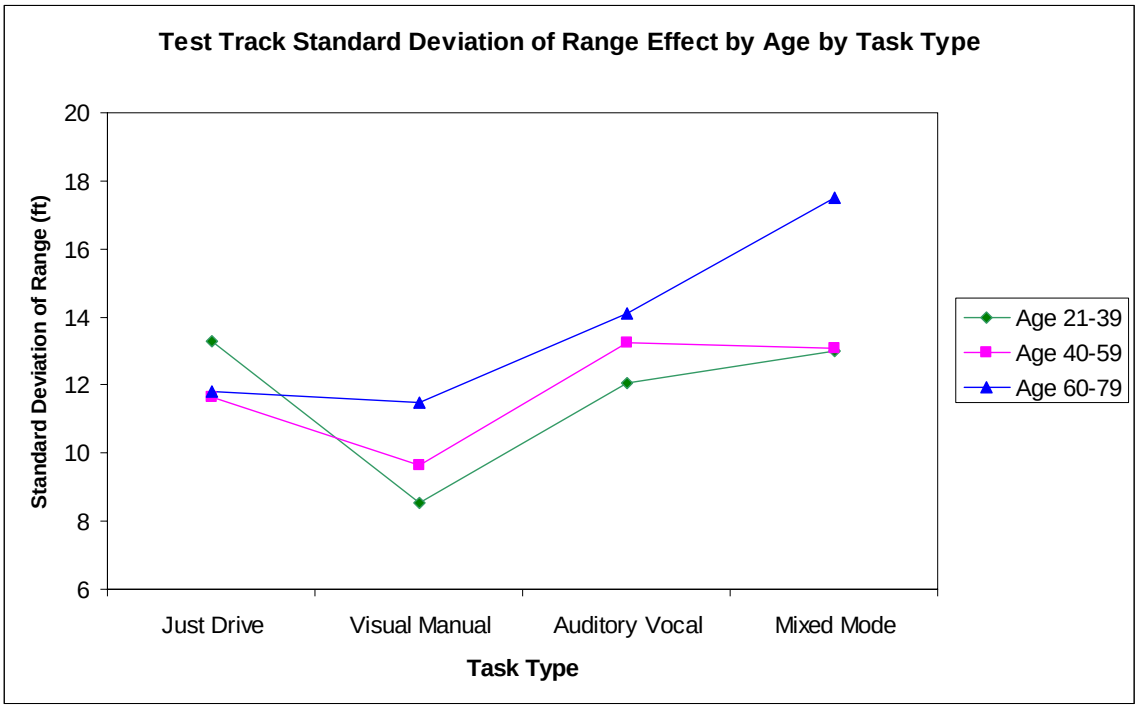


Figure 7-18. Age Effect – Test Track Standard Deviation of Range

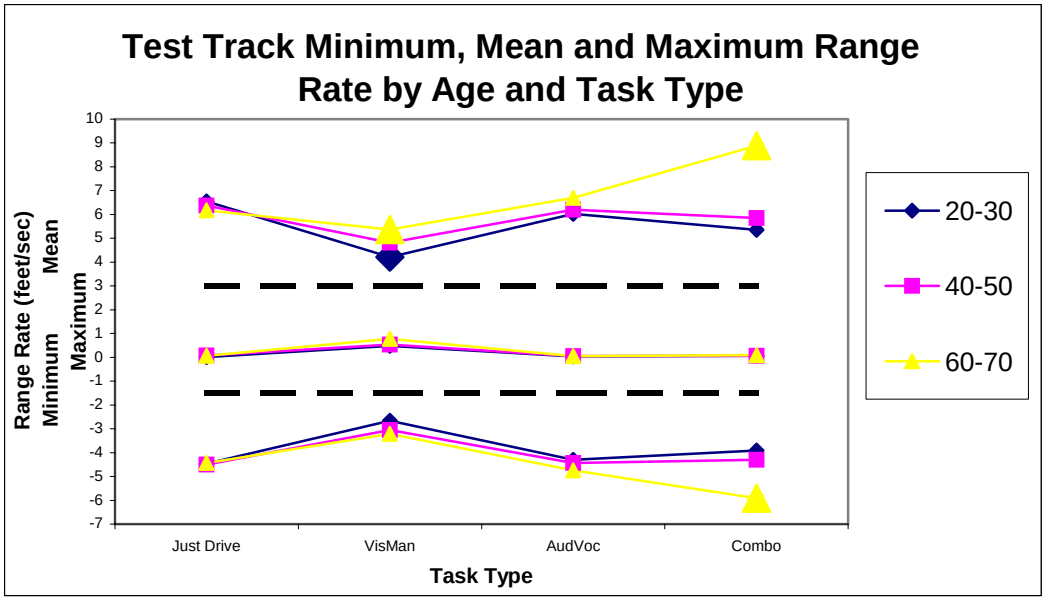


Figure 7-19. Age Effect – Test Track Range Rate

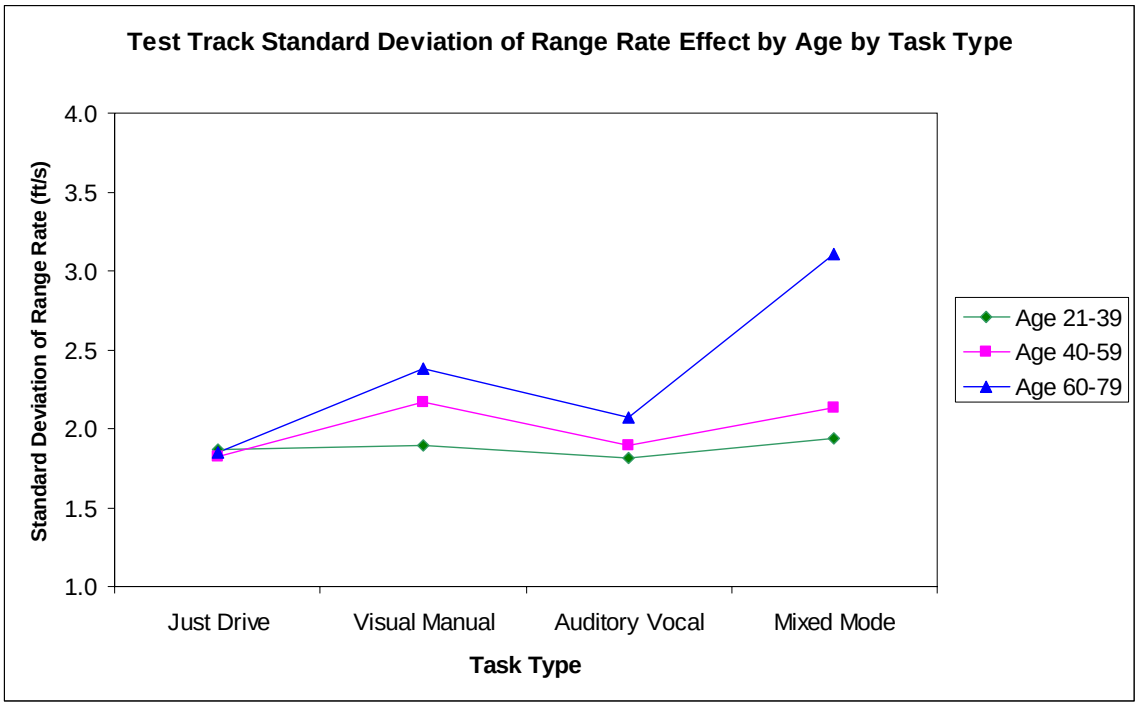


Figure 7-20. Age Effect – Test Track Standard Deviation of Range Rate

Figure 7-21 presents the age and task effects for minimum, mean, and maximum speed by task type and age group. Minimum speed shows a significant age by task interaction with a difference between younger and older test participants for the visual-manual task type, but not for the mixed-mode task type. While there is a significant task effect on mean speed, there is not a significant age effect. For maximum speed, there is an effect of age for three task types, inverse of the age effect with minimum speed. For maximum speed, there are differences with all but the visual-manual task type. Older test participants still exhibit more variability with the mixed-mode task as well as with Just Drive and auditory-vocal task type. The absolute differences in speed are quite small, so practical significance must be determined, but there is an effect due to age group.

Figure 7-22 shows the task and age effects on standard deviation of speed. There is a significant difference between older and younger test participants for the task types with a manual component. The trend of significantly higher variability for older test participants with these shorter duration task types was seen in many of the test track results.

Figure 7-23 shows significant task and age effects for Speed Difference, which is calculated as the maximum speed during a task minus the minimum speed. For all age groups, the visual-manual task type is the only type that differed significantly from the other task types. However, the similarity of the shorter mixed-mode task with the longer Just Drive and auditory-vocal task type shows that there is more variability in speed in a shorter duration for the mixed-mode task type. It is interesting to note also that for older test participants the between tasks trend differs from the other age groups. With the two other age groups, the mixed-mode task has a slightly lower speed difference whereas for older test participants, the speed difference is greater. For all task types except Just Drive, there is a significant effect of age, with the older test participants demonstrating more variability in speed than the other age groups.

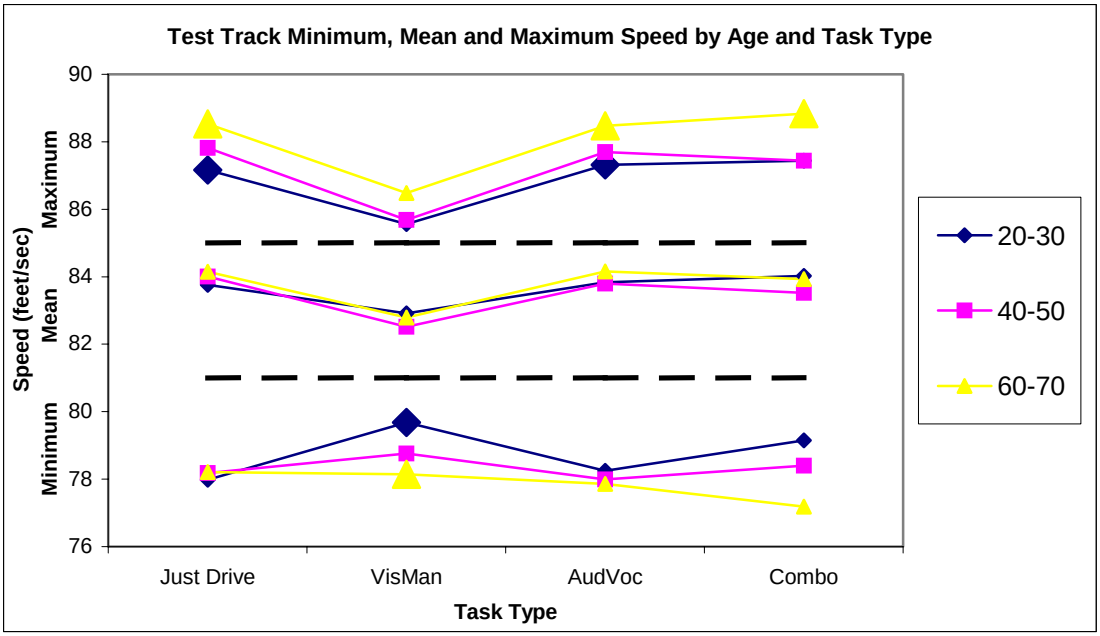


Figure 7-21. Age Effect – Test Track Speed

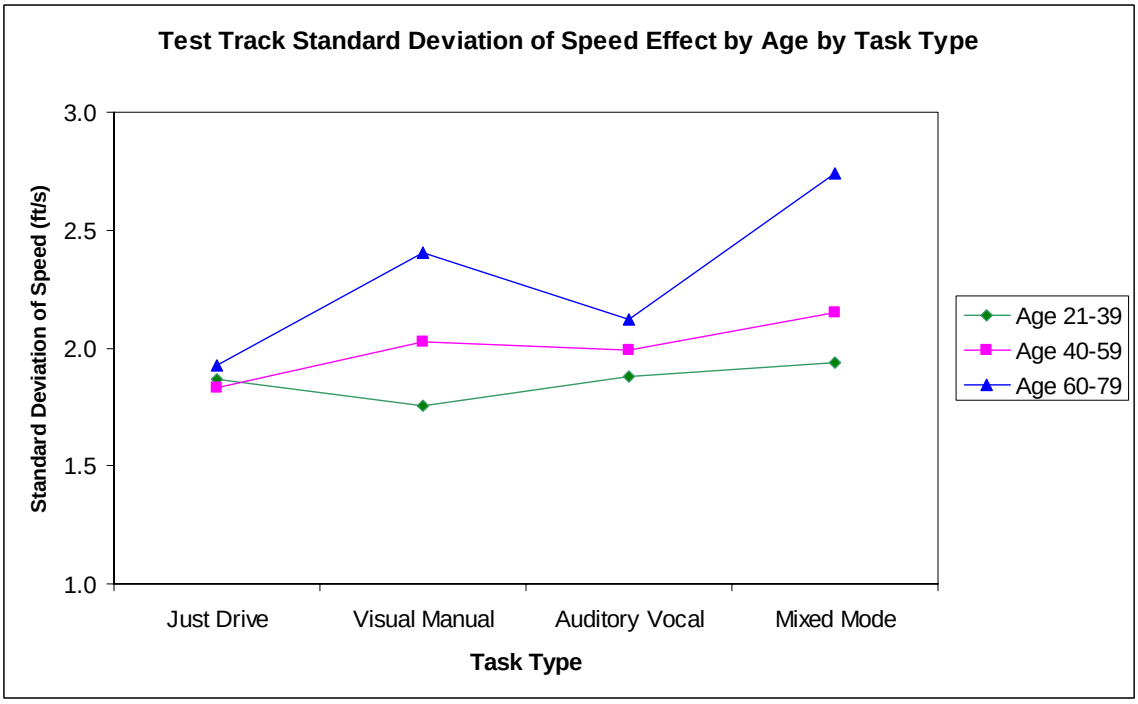


Figure 7-22. Age Effect – Test Track Standard Deviation of Speed

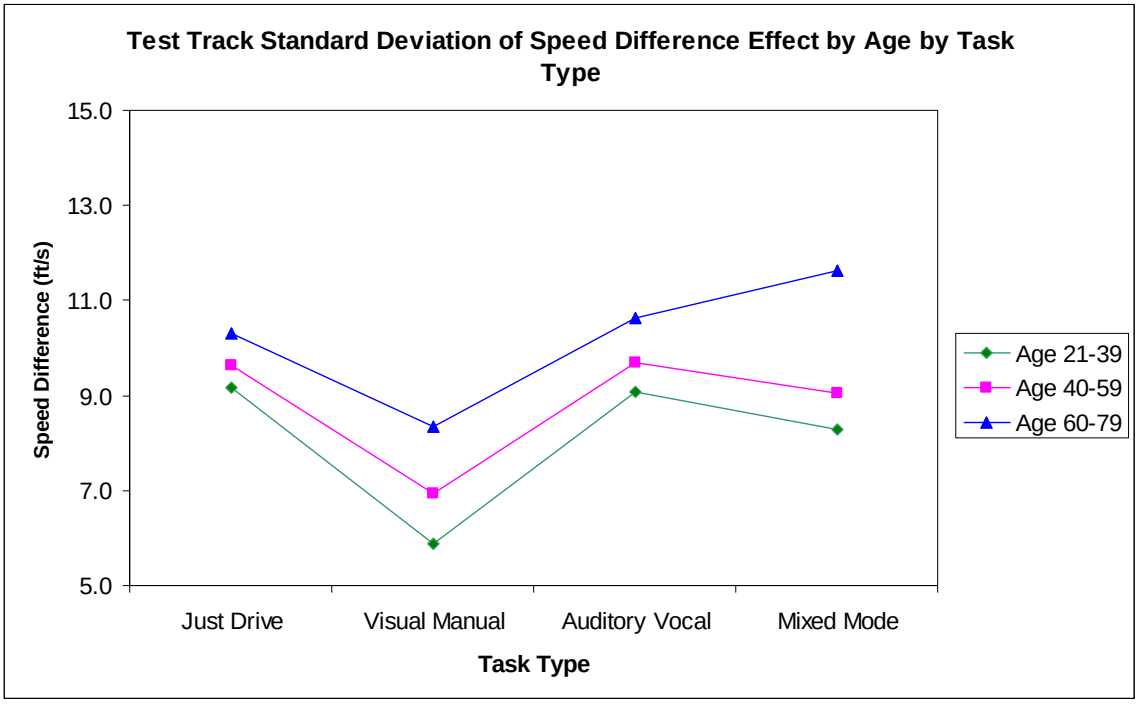


Figure 7-23. Age Effect – Test Track Speed Difference

SDLP shows significant age effects with the oldest test participants having the most variation in lane position. There is also a task-by-age interaction for Just Drive and the auditory-vocal task type. In these cases, the middle-age group has the tightest lateral control while for the manual-component tasks the youngest test participants perform better.

For this venue, longitudinal measures are very similar to the on-road data. There are differences, due in part to the additional tasks performed on the track, which lead to age-by-task interactions. Typically however, the oldest test participants have the most variation in longitudinal control, especially for the manual component tasks.

7.4 Gender Effects

There are recognized differences between male and female drivers as evidenced by both accident statistics and insurance rates. Males and females are also likely to perform in-vehicle tasks differently as evidenced by differences in various cognitive test performances. With the development of new types of in-vehicle devices, understanding their potential effects on both males and females is critical. This section examines the effects of gender on surrogate and driving performance metrics.

7.4.1 Laboratory Gender Effects

Figure 7-24 shows the effect of gender on TSOT. While static task times are almost identical for the two groups, TSOT shows an effect for the mixed-mode task type where male test participants have longer task times than do female test participants indicating a task by gender interaction.

Figure 7-25 shows the gender effect on STISIM SDLP. The gender groups are nearly equal for the mixed-mode task type, but males have slightly tighter lateral control for Just Drive and auditory-vocal task types. Both genders have significantly higher SDLP for the visual-manual

task type with males having the most variation in lateral control for this task type. This trend is in striking contrast with that seen later for on road and test track measures of SDLP.

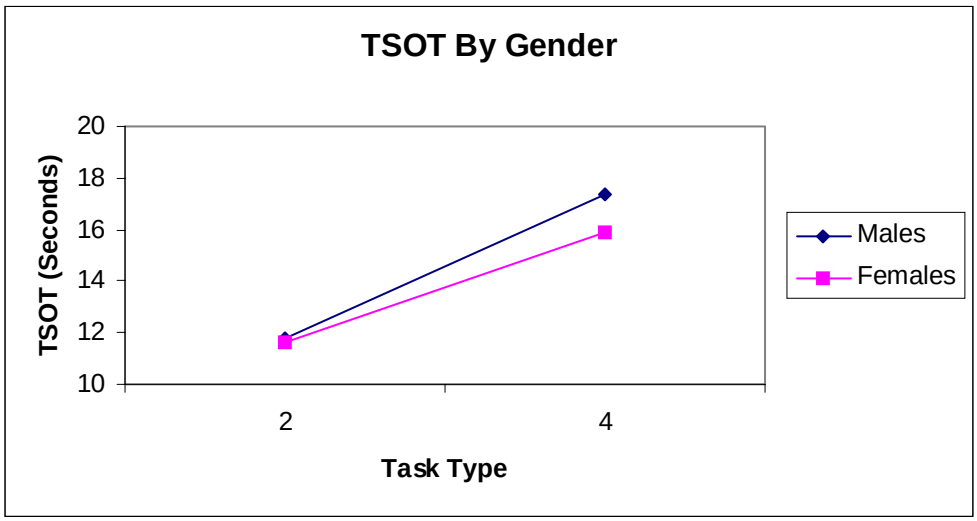


Figure 7-24. Gender Effect – Laboratory TSOT by Task Type

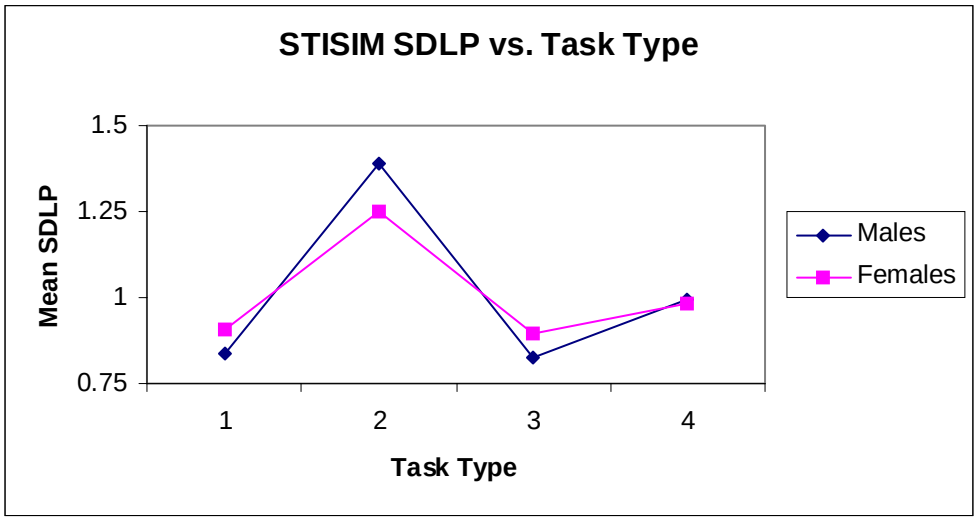


Figure 7-25. Gender Effect – STISIM SDLP by Task Type

The mean speed difference by task for the STISIM surrogate is presented in Figure 7-26. For this measure, both genders show task type 2 levels similar to those for Just Drive while the auditory-vocal task type is higher and the mixed-mode task type is the highest. The female test participants show significantly and uniformly more difference in speed than the male test participants.

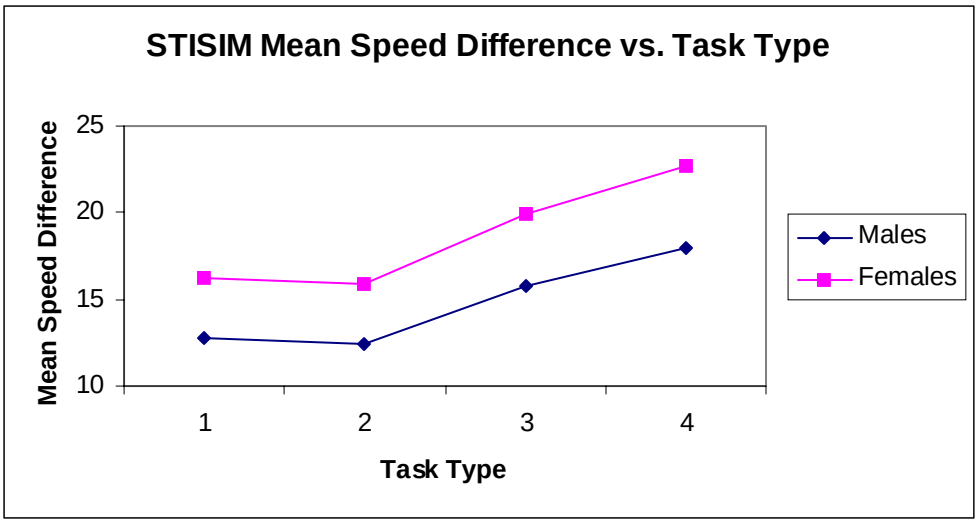


Figure 7-26. Gender Effect - STISIM Speed Difference by Task Type

The miss rate for PDT-Alone for each gender and task is shown in Figure 7-27. Both genders show higher miss rates for the visual-manual and mixed-mode task types. Male test participants, however, have higher miss rates for all but the Just Drive task and this difference is greatest for the mixed-mode task. This trend is also seen in Figure 7-28 for PDT with STISIM, except the increase in difference between the genders is not present for the mixed-mode task. This indicates a possible interaction between task, gender, and the surrogate.

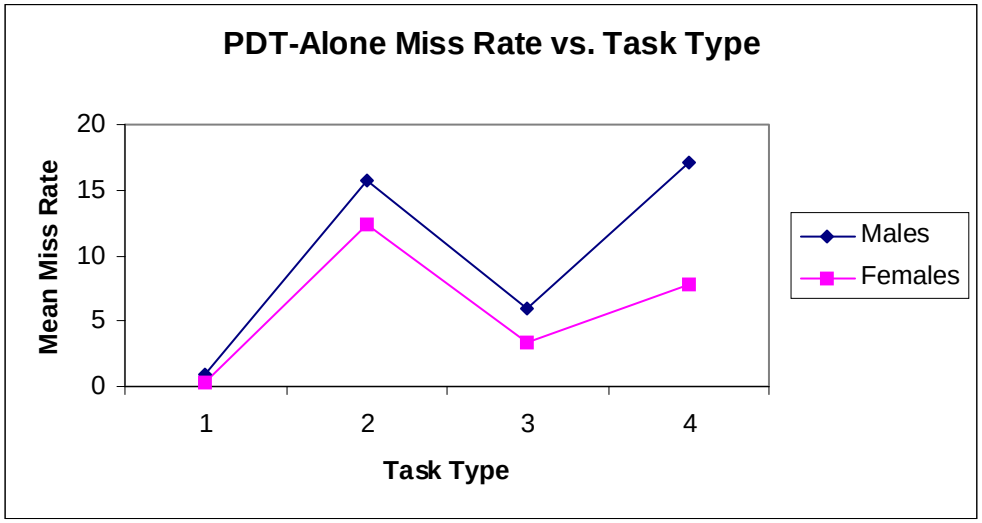


Figure 7-27. Gender Effect – PDT-Alone Miss Rate by Task Type

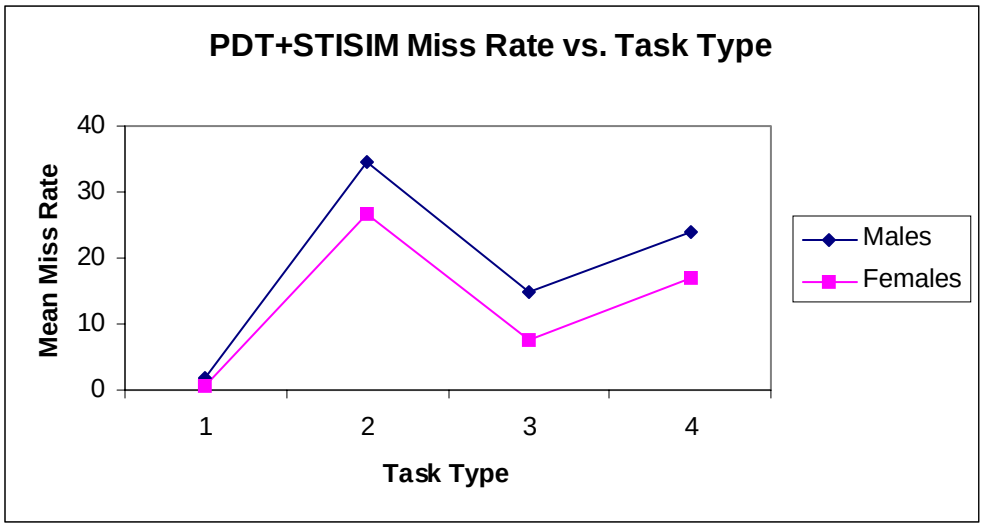


Figure 7-28. Gender Effect – PDT With STISIM Miss Rate by Task Type

The Sternberg memory task percent error results by gender are presented in Figure 7-29. For this surrogate, there is a trend, both between tasks and between genders, very similar to that for PDT with STISIM miss rates. Again, male test participants have higher miss rates for all but the Just Drive task type with a slight increase in the difference between genders when it comes to the mixed-mode task type.

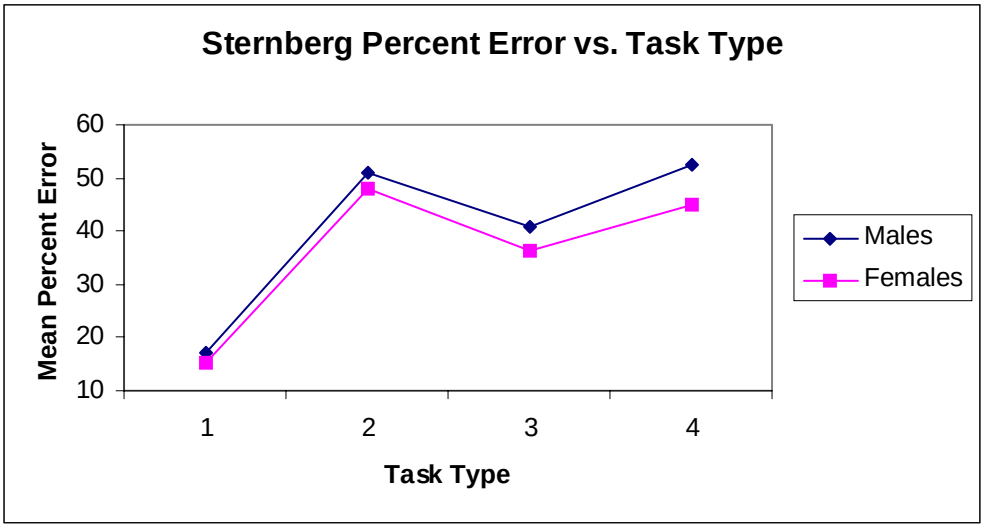


Figure 7-29. Gender Effect – Sternberg Percent Error by Task Type

Patsys Grammatical Reasoning Tests have two notable gender effects, as shown in Figure 7-30. The scale for the average number of correct responses, in this figure, is shown on the auxiliary y-axis on the ride side of the figure. This graph shows the difference between genders for both the number correct and percent correct. The pattern between both measures is the same, with females having higher scores for both metrics. This indicates that not only are the female test participants

performing the test faster and answering more questions, they are more accurate in their responses than their male counterparts.

Useful Field of View scores can be seen in Figure 7-31 for processing speed, selective attention, and divided attention. While the differences in processing speeds between genders are small, on average, female test participants are scoring 25 percent better for this measure. For divided attention, females again score better than males, with an approximately 50 percent better score. Selective attention shows the same pattern with female test participants scoring approximately 25 percent better than their male counterparts.

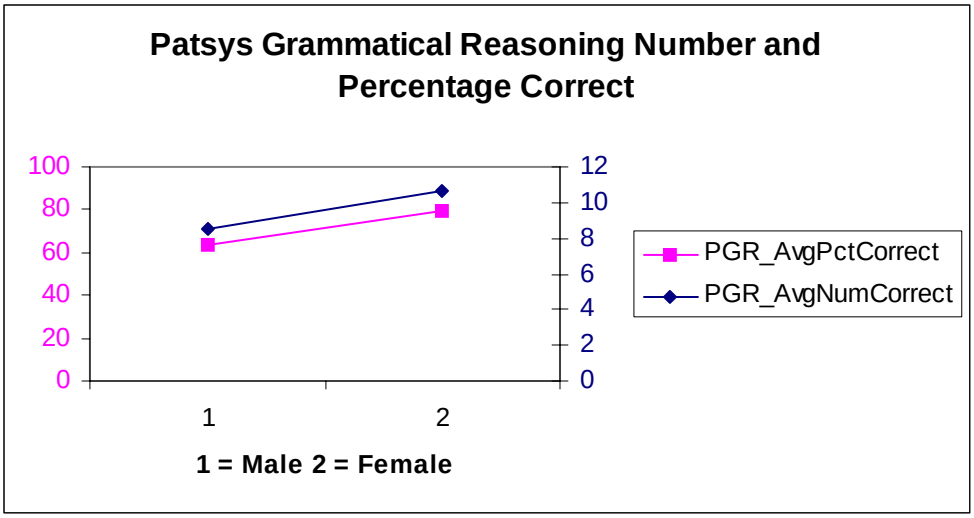


Figure 7-30. Gender Effect – Patsys Grammatical Reasoning

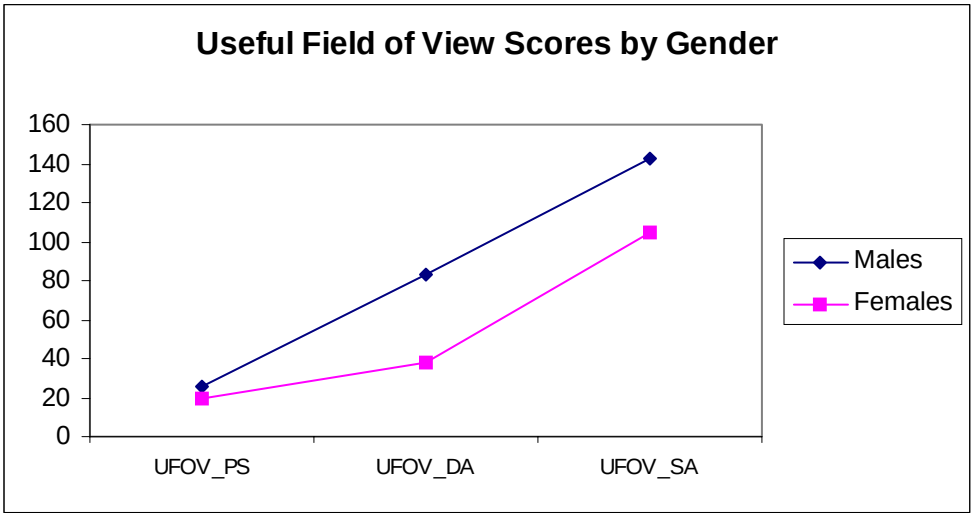


Figure 7-31. Gender Effect – Useful Field of View

While static task times were nearly equal for the gender categories, male test participants had a slightly higher TSOT and thus lower R-Metric for the mixed-mode task, a difference that is not observed in the visual-manual task type.

The STISIM lateral control measures had mixed differences between the genders indicating, in some cases, possible task by gender interactions. Absolute differences are, however, relatively small. STISIM longitudinal control in the form of speed difference indicates a large difference between the genders with female test participants having a relatively uniformly larger speed difference for all tasks. PDT-Alone and PDT with STISIM both show that male test participants have a higher miss rate for all task types except Just Drive. PDT with STISIM shows approximately twice the absolute difference between the genders that PDT-Alone does. PDT with STISIM, however, has a higher miss rate in general so the difference is not quite that large when compared to the higher miss rates.

The Sternberg memory test exhibits a trend for male test participants to have a higher error rate for tasks than do females. The difference seems to indicate a gender by task interaction. For Patsys Grammatical Reasoning, female test participants answer about 20 percent more questions correctly and are answering a greater percentage of their questions correctly. For the Useful Field of View Measures Selected and Divided Attention female test participants are again scoring significantly better than the male test participants.

7.4.2 On-Road Gender Effects

Figure 7-32 shows task and gender effects for SDLP. There is only a significant difference for the mixed-mode task, with female test participants having a slightly higher SDLP. Female test participants also exhibit a break with male test participants in the between task trend with the mixed-mode task. Here female test participants show more variability rather than less when compared to the longer auditory-vocal task type.

Figure 7-33 shows task and gender effects for maximum range, which has been shown in other graphs to be sensitive to task and individual differences effects. In this instance, the significant difference is that female test participants appear to be falling back more during the auditory-vocal task type.

Figure 7-34 shows the task and gender effects for the standard deviation of range, a measure of overall variability in range. For all except the visual-manual task type, female test participants tend to have higher variability in this longitudinal measure.

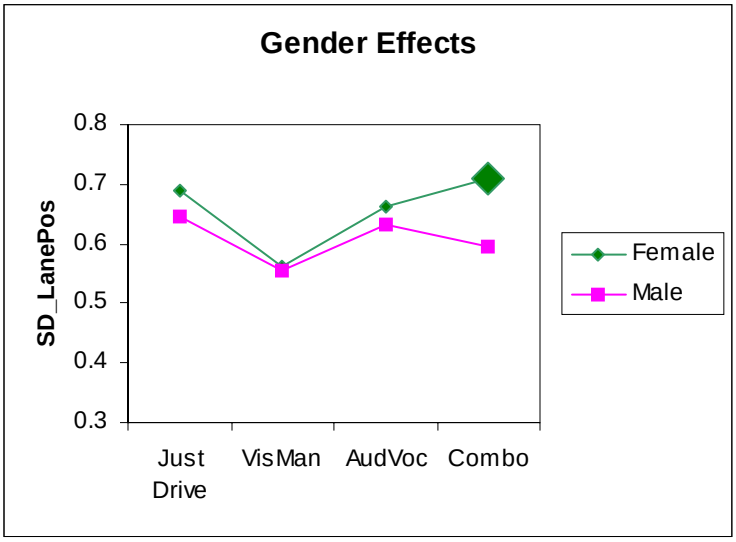


Figure 7-32. Gender Effect – On-Road Standard Deviation of Lane Position

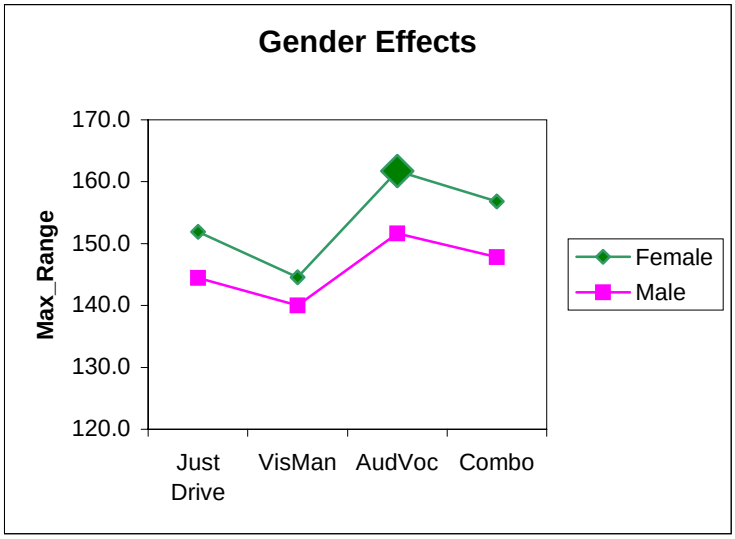


Figure 7-33. Gender Effect – On-Road Maximum Range

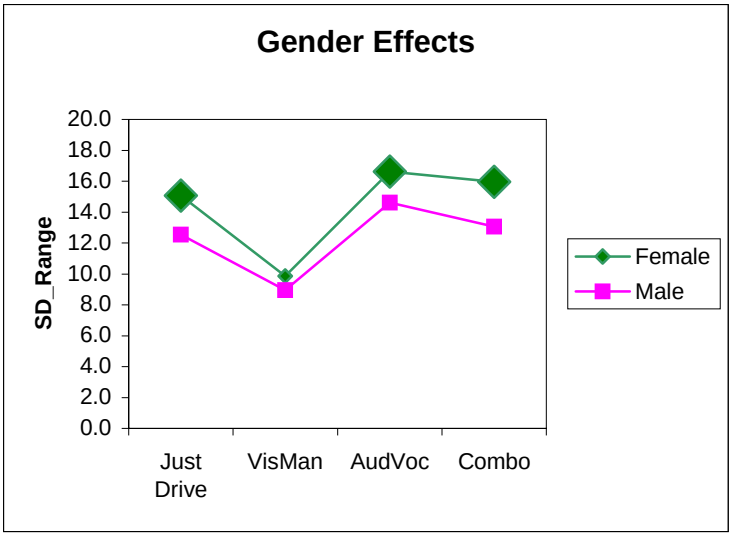


Figure 7-34. Gender Effect – On-Road Standard Deviation of Range

Figure 7-35 presents the task and gender effects for minimum, mean, and maximum range rate. Minimum range rate shows a significant difference only for the mixed-mode task type, while mean range rate has no statistically significant differences. Maximum range rate is again the most significant measure showing significant differences for all but the visual-manual task type, similar to the results for standard deviation of range.

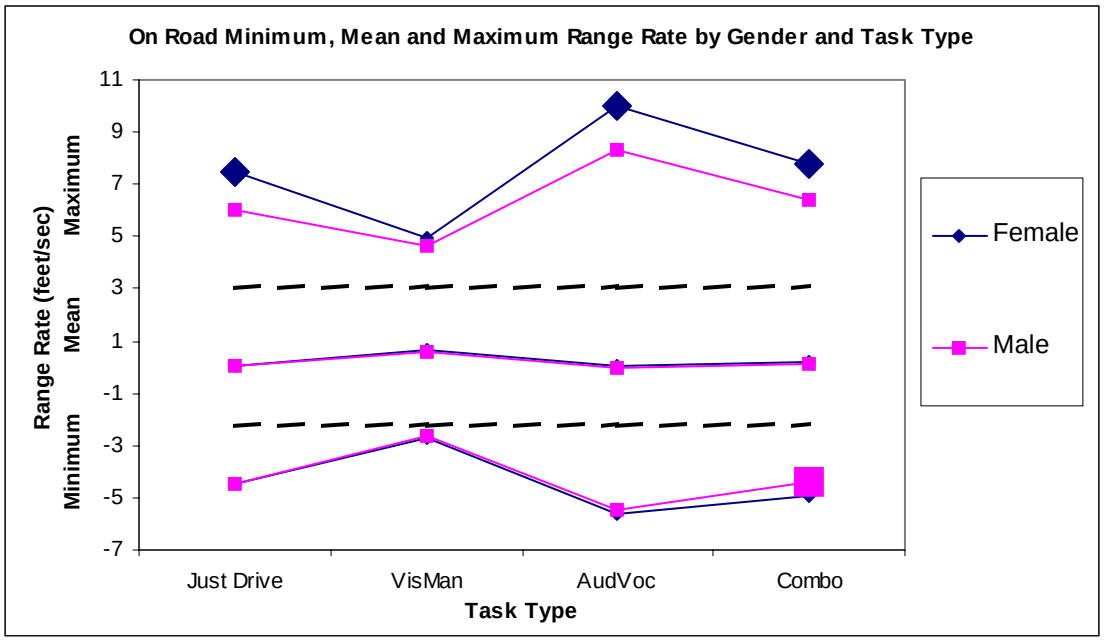


Figure 7-35. Gender Effect – On-Road Range Rate

Figure 7-36 shows the task and gender effects for standard deviation of range rate. Again there are significant differences in the means for all but the visual-manual task type. While measures such as minimum and maximum range rate show lower values for the mixed-mode task than for

auditory-vocal tasks, the measure here indicates that the mixed-mode task type has greater variation for both genders.

Figure 7-37 presents the task and gender effects for minimum speed. It can be seen that male test participants have a higher minimum speed for tasks with significant differences for the Just Drive and mixed-mode task types. This result is expected given the lower variation in range statistics seen in previous graphs.

Figure 7-38 shows the task and gender effects for standard deviation of speed. Both genders have as much variation in speed during the short visual-manual task type as in the longest, Just Drive task type. Variation increases during the auditory-vocal task type with the most variation coming during the mixed-mode task type. The mixed-mode task type also shows a significant difference between male and female test participants in speed variation, though again, the absolute difference is rather small.

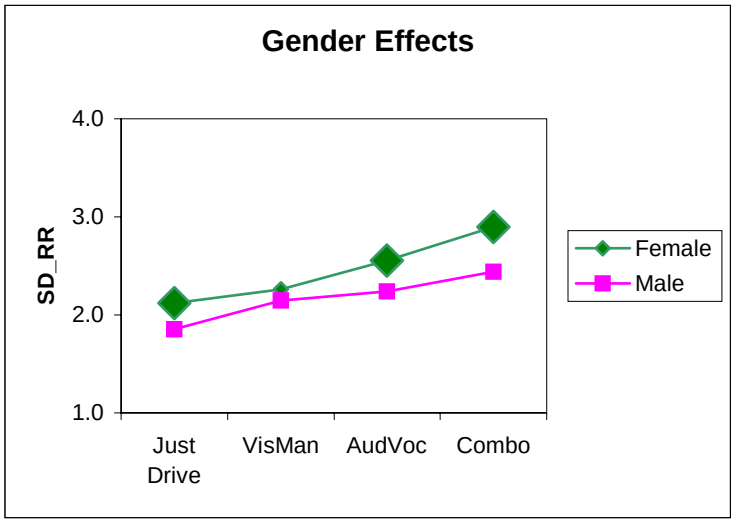


Figure 7-36. Gender Effect – On-Road Standard Deviation of Range Rate

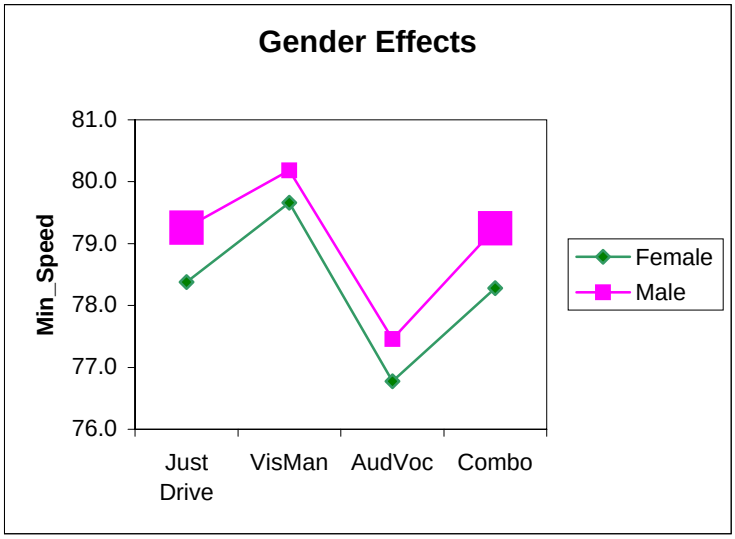


Figure 7-37. Gender Effect – On-Road Minimum Speed

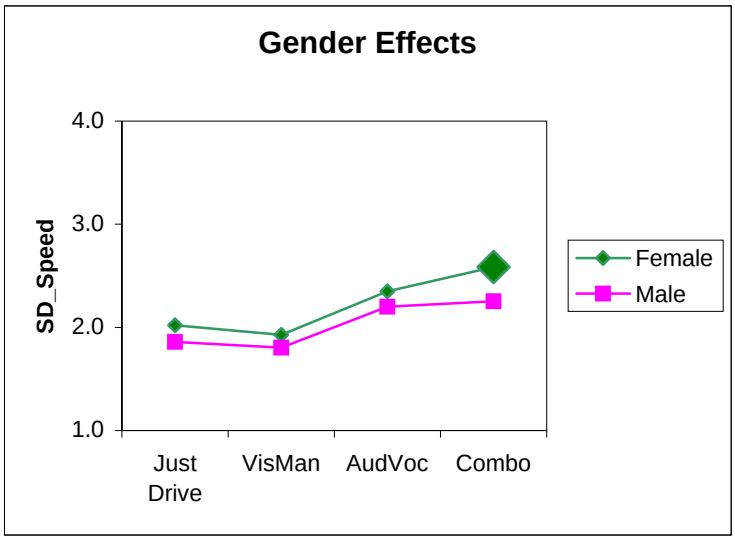


Figure 7-38. Gender Effect – On-Road Standard Deviation of Speed

Figure 7-39 presents the task and gender effects for speed difference, defined as maximum minus minimum speed. Here it can be seen that visual-manual tasks have the lowest difference and auditory-vocal tasks have the most change. With the exception of Just Drive, the speed difference follows task duration, with the mixed-mode task type in the middle for both genders. Throughout this pattern, female test participants have higher speed differences with the significant differences being for Just Drive and mixed-mode task types.

For test track SDLP, female test participants have a significantly higher value for the Voice Dial mixed-mode task. This is due to a break in the trend of other tasks where the two genders have nearly equal variation.

Female test participants have significantly higher longitudinal-control variation means for a number of tasks, most often auditory-vocal and mixed-mode task types. Female test participants also tend to have lower minimum speeds, higher maximum-range rates, and higher speed differences than male test participants, which indicate female test participants are slowing more during tasks than male test participants. These differences are relatively small and practical significance is a question to address.

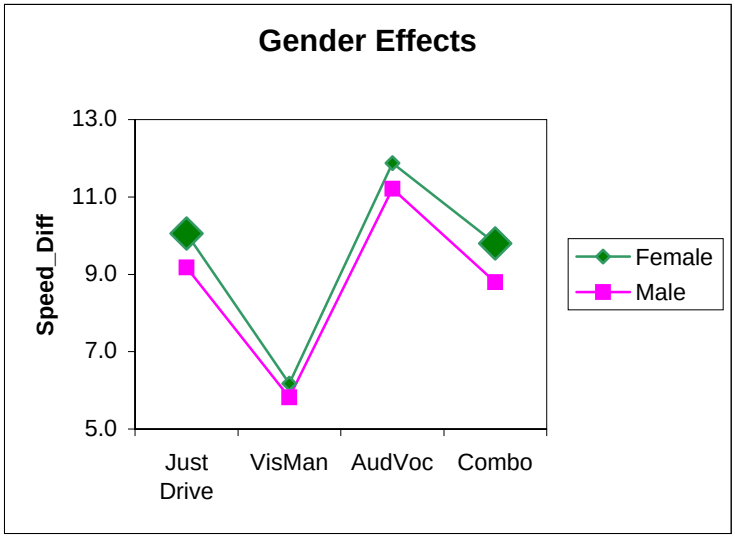


Figure 7-39. Gender Effect – On-Road Speed Difference

7.4.3 Test Track Gender Effects

Figure 7-40 presents the task and gender effects on total number of lane exceedances. Unlike the on-road results, where females had roughly equal number of lane exceedances for all but the mixed-mode task type, male test participants have the higher number for the test track. This measure does not have a significant gender effect for on-road or test track, however with the change in venue and additional tasks, there is a statistically significant difference. The difference comes with visual manual tasks where male test participants have roughly double the lane exceedances.

Figure 7-41 shows the task and gender effects for minimum, mean, and maximum range. While minimum range does not show a significant effect of gender, there is a significant difference between males and females on the auditory-vocal task type, with male test participants having a higher minimum range. Mean range has a significant task effect and age by gender interaction, but no age or gender effects. Maximum range shows significant effects for task, age, gender, age-by-gender and task-by-age. As the graph shows, female test participants have higher maximum ranges for the visual-manual and mixed-mode task types. Interestingly, whereas male test participants have a maximum range that is lower for the mixed-mode task than that for the auditory-vocal task type, female test participants exhibit the opposite trend.

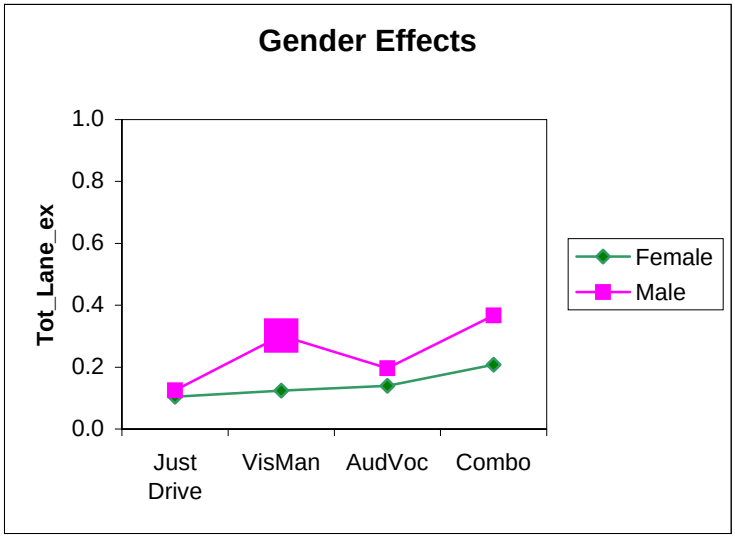


Figure 7-40. Gender Effect – Test Track Total Lane Exceedance

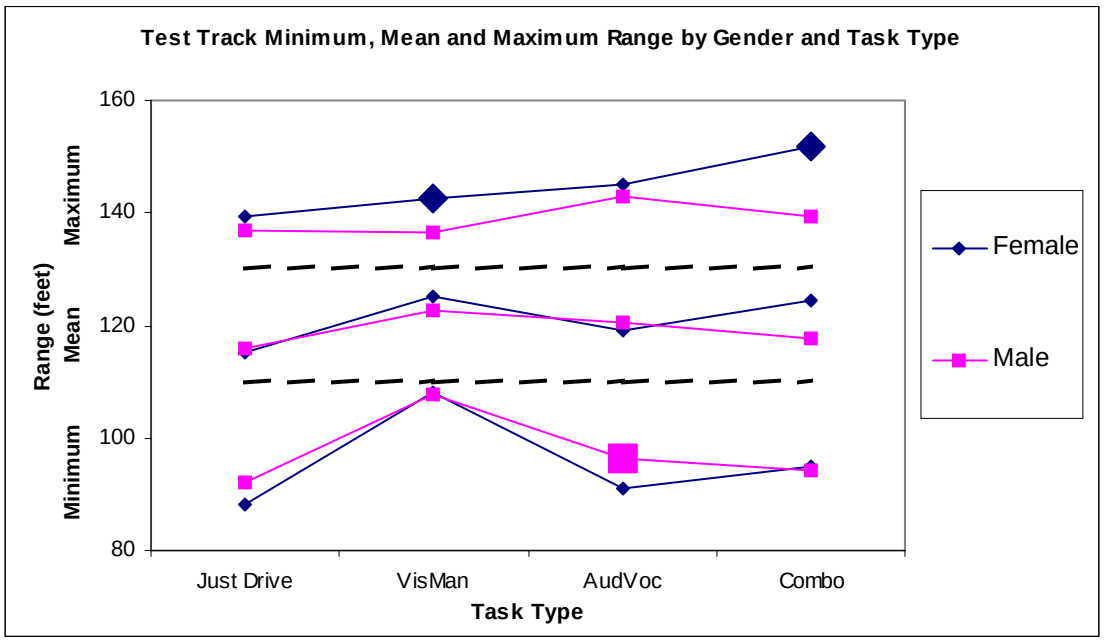


Figure 7-41. Gender Effect – Test Track Range

Figure 7-42 shows the significant task and gender effects on standard deviation of range, where there are also significant age-by-gender and task-by-age interactions. The gender effect is interesting here because of significant differences for all four task types. The Just Drive and visual-manual task types show a very nearly equal difference between male and female test participants. For the auditory-vocal task type, this difference increases by about 25 percent. The fourth task type, the mixed-mode type, stands out yet again with two differences in the trend between tasks. Male test participants have a standard deviation of range for the mixed-mode task that is nearly equal to the auditory-vocal task type. Female test participants show a difference between auditory-vocal and mixed-mode task types that is four times larger than the difference for male test participants. This also breaks the uniformity between the genders that is seen with the other tasks, with nearly double the difference between genders than that seen with the first two task types.

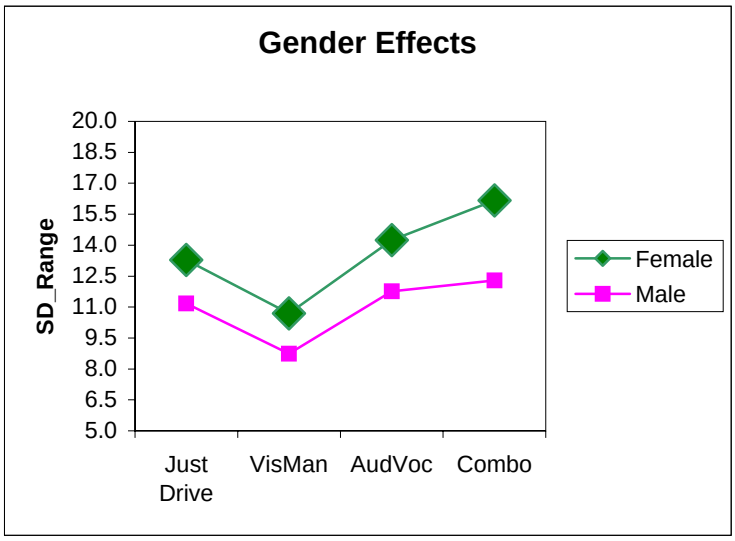


Figure 7-42. Gender Effect – Test Track Standard Deviation of Range

For this venue, females once again exhibit more variation in lateral and longitudinal control with higher lane exceedance counts for visual-manual tasks as well as higher maximum range and higher standard deviation of range.

7.5 Predictive Power of Cognitive Tests

It is clear that cognitive abilities of drivers affect their driving performance. Understanding these differences can be key to predicting driving behavior and ability to multitask while driving. Understanding how these differences affect driving can also aid in product design and development. If these tests can predict driving performance then they can be used to estimate subsidiary in-vehicle tasks' affect on driving, and give designers an indication of tasks that require refinement. .

7.5.1 Useful Field of View (UFOV)

Useful Field of View is defined as the size of an area from which a person can obtain and process information with only a brief glance and no head or eye movement. Visual Awareness Inc. developed tests based on this concept to measure the time an individual requires to gather information.

The first of these tests is called Processing Speed. In this test, a test participant focuses on a fixation box presented on a computer screen. Periodically, an image of a car or truck is displayed for a decreasing duration in the fixation box. The score of this test is the minimum time in milliseconds that a test participant was required to correctly identify the object that was presented.

The second test is called Divided Attention and it measures a test participant's ability to process two images. With this test, a target, car or truck, is again presented in the fixation box. Simultaneously, the silhouette of a car is also presented along one of eight radial axes. The test participant must identify what the image in the box was as well as where the second object appeared on the screen. The score of this test is the minimum time in milliseconds a test participant requires to correctly identify both the central target as well as the location of the second target.

The last test in this group is called Selective Attention. The task is the same as in Divided Attention, however, now all the possible locations of the secondary target that are not occupied by the target are filled with triangles. These distracters require the test participant to focus enough attention on the secondary target to differentiate it from the distracter triangles. The score again is the minimum time in milliseconds it takes a test participant to correctly identify the primary target and location of the secondary target.

7.5.2 Selective Attention

For this measure, an analysis of variance shows significant effects with selective attention recoded based on the upper and lower-thirds of the distribution. In the legends for the charts presented in this section (See Figure 7-43, Figure 7-44, Figure 7-45, and Figure 7-46), the lower third of the data distribution is labeled one (1), the middle is labeled two (2), and the higher third is labeled three (3). As with other cognitive tests, some measures show rather little variation between test participants. For on-road data, selective attention has a significant effect on standard deviation of lane position, standard deviation of speed, and speed difference. Figure 7-43 shows this effect for the mean standard deviation of lane position, with only the visual-manual task type showing a difference between groups. Figure 7-44 shows the difference between selective attention categories with respect to the mean standard deviation of speed, again the largest differences between groups is seen with the tasks having a visual-manual component. Figure 7-45 shows that each of the three groups has similar performance in speed difference. Each of the groups has the familiar trend between tasks: longer tasks showed more variation than the shorter tasks.

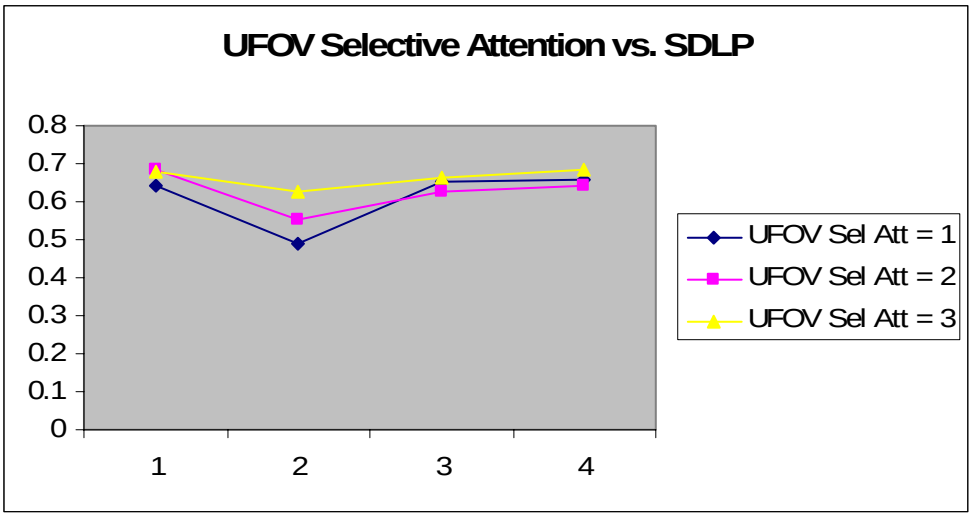


Figure 7-43. UFOV Selective Attention Versus On-Road SDLP

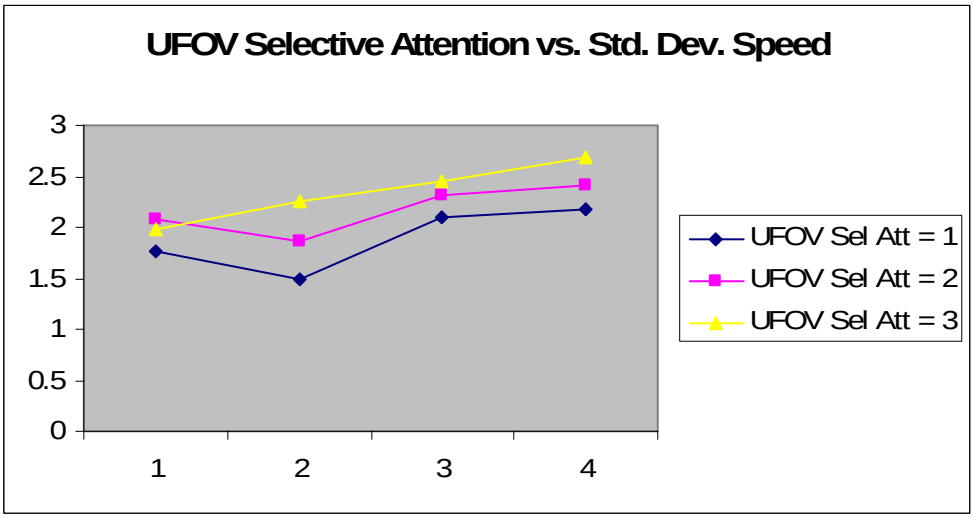


Figure 7-44. UFOV Selective Attention Versus On-Road Standard Deviation of Speed

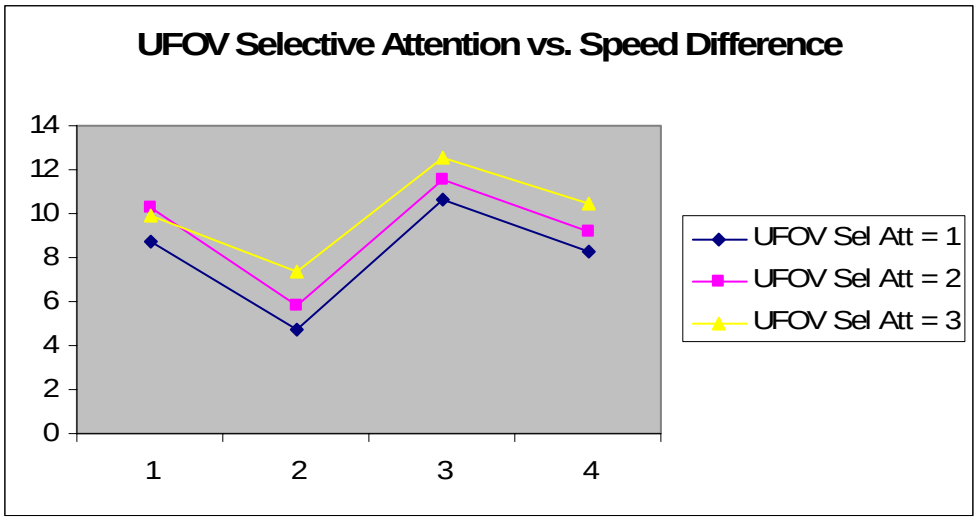


Figure 7-45. UFOV Selective Attention Versus On-Road Speed Difference

Figure 7-46 shows the connection between selective attention and the standard deviation of speed for the test track participants. Generally, this measure shows the same patterns between road and track. For the standard deviation of speed measure for the test track participants however, the addition of tasks to types 2 and 4 caused a change in the relative locations between groups. The best performers have a rather flat trend between tasks while the other two groups show more variability in speed for the visual-manual type tasks.

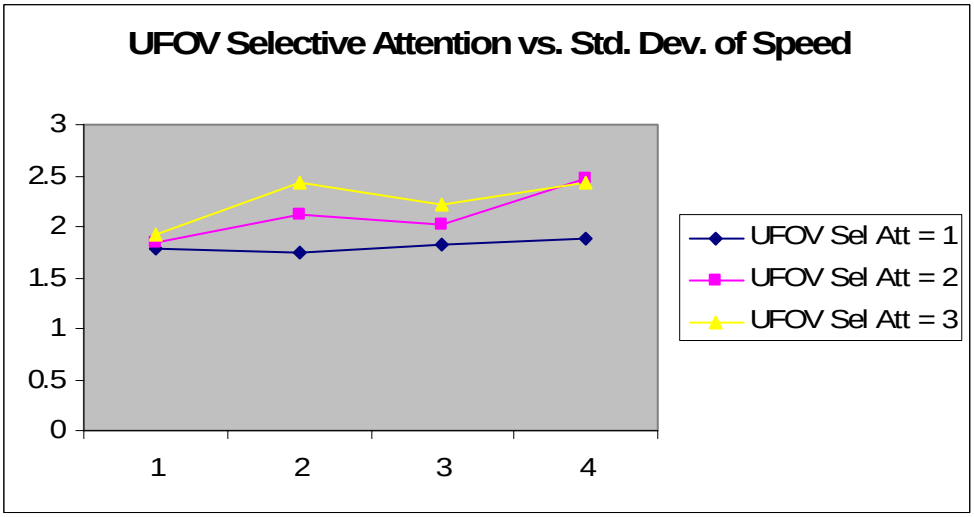


Figure 7-46. UFOV Selective Attention Versus Test Track Standard Deviation of Speed

7.5.3 Divided Attention

Divided Attention for both test track and on road participants was categorized based on upper and lower halves of the distribution. In the legends for the charts presented in this section (See Figure 7-47, Figure 7-48, Figure 7-49) the data was categorized into a lower group labeled one (1) and a

higher group labeled two (2). Divided attention, like selective attention, showed significant effects for three measures of driving performance. Figure 7-47 shows the difference between divided attention groups for test track mean SDLP. Group 2, with the lower performance on the test, shows less between task variability. Figure 7-48 shows that group 2 has more variation in on-road speed for all but the Just Drive task type. The two groups also differ in between task patterns with opposite trends for visual-manual and mixed-mode task types. Figure 7-49 shows the mean on-road speed difference for the two groups of divided attention scores. As seen previously, the lower performing group, Group 2, also has more variation in this vehicle control metric. The between task trend for both groups roughly follows mean task duration for all but the Just Drive task type.

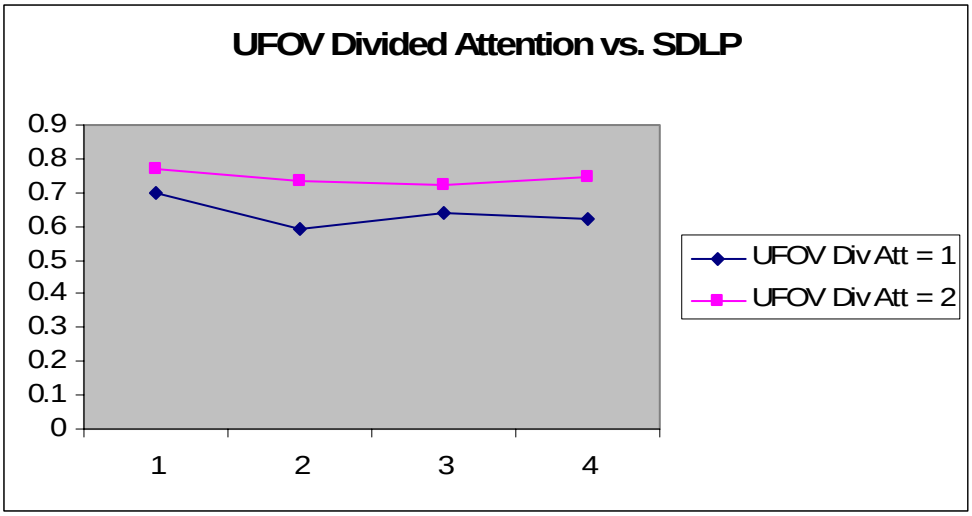


Figure 7-47. UFOV Divided Attention Versus Test Track Mean SDLP

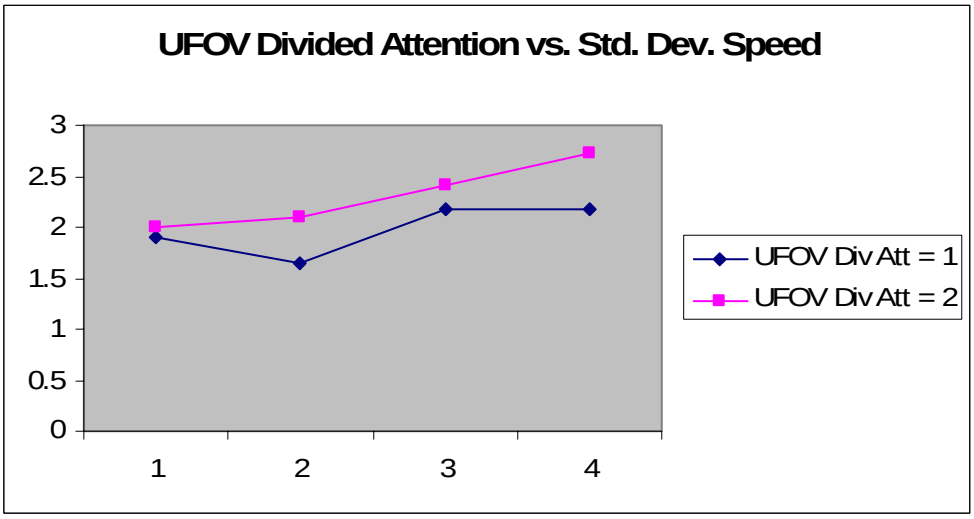


Figure 7-48. UFOV Divided Attention Versus On-Road Standard Deviation of Speed

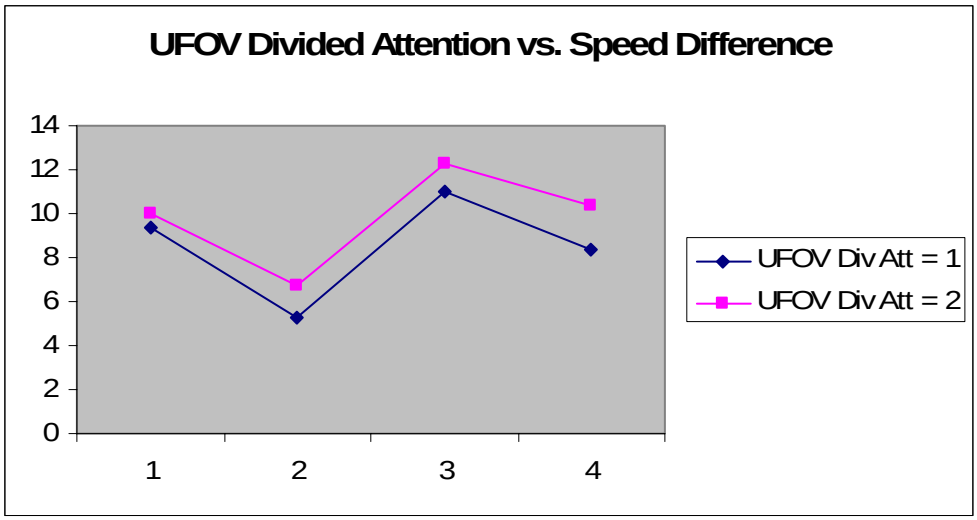


Figure 7-49. UFOV Divided Attention Versus On Road Mean Speed Difference

UFOV Selective Attention has significant effects for SDLP, standard deviation of speed, and speed difference. For SDLP, only the visual-manual task type has significantly different means between levels of Selective Attention (SA). Standard deviation of speed is more varied between the levels for all tasks but the differences are still relatively small. The different levels of SA are more clearly separated when examining speed difference, however, the trend roughly follows task duration and absolute differences are relatively small.

For the test track, results are very similar but possibly show significant task by selective attention interactions that do not seem to be present with on-road data. For both venues, R-squared values are very small, so the test will only rank order potential performance.

UFOV Divided Attention also has significant effects for SDLP, standard deviation of speed, and speed difference, with the results indicating the same trends as with selective attention for both on-road and test track.

7.5.4 Patsys Manikin (PM)

Patsys Manikin is a computer-based test used to evaluate a test participant's ability to process visual information. This test battery consists of three individual tests.

The first test, Patsys Temporal Acuity, has a black box in each corner of a computer monitor. Periodically a white image, similar to a Langholdt C, is displayed in each box sequentially, facing either right or left. The test participant must focus on the left-hand box and then shift focus to the right-hand box after seeing the probe appear in the left-hand box. The test participant is instructed to answer "same" or "different" to indicate whether or not the displayed images are oriented in the same direction or different directions.

The second test in this battery, Patsys Manikin, briefly displays a simple outline drawing of a sailor on a computer screen. The sailor is standing on a box-shaped platform containing images of diamonds, hearts, spades, or clubs. A similar box, each containing different images of the four card suits, is also held in each of the sailor's hands. Randomly, the sailor image will be facing toward or away from the test participant. The test participant is then to indicate, via computer keyboard arrow keys, whether the box that matches the platform is in the sailor's left or right hand.

The third test in this battery is Patsys Grammatical Reasoning. In this test, text is displayed on the computer screen. A short statement, such as “A does not follow B”, will appear next to two letters such as “AB”. The test participant’s response is “true” or “false”. In this example, the correct response is true, indicated by pressing the “T” key on a computer keyboard.

Results for all tests are number of probes, number of correct responses, and reaction times. Reaction time values are the mean reaction time for all correct responses. The percentage of correct responses (percent correct) was calculated from the number of probes and number of correct responses and subsequently analyzed.

7.5.4.1 PM Percent Correct

Patsys Manikin percent correct has a significant effect with on road SDLP as can be seen in Figure 7-50. Here, the group with the lowest percentage of correct responses also has the highest SDLP. Interestingly, the greatest variation between groups is seen with the Just Drive task, and there is little difference in SDLP between the two groups with the best test scores.

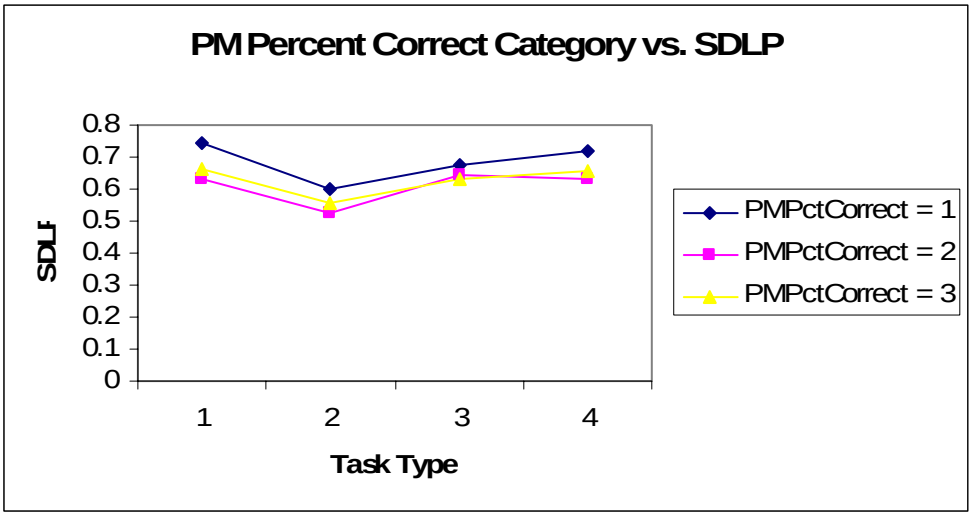


Figure 7-50. Patsys Percent Correct Versus On-Road Mean SDLP

For the test track participants, the distribution of scores for Patsys Manikin percent correct warranted only splitting the test participants into two groups. The difference in SDLP for these two groups can be seen in Figure 7-51. While there is not a statistically significant effect here, it is interesting to note a few changes in the plot when compared with the on-road test participants. Just Drive again shows the lowest test score group having the higher SDLP, and both groups have higher SDLP for this than other tasks. Unlike on-road test participants, with this group the lower test-score group had slightly better SDLP for the visual-manual and mixed-mode task types. Overall however, the between groups differences are very small. The distribution of scores split test participants into two groups causing a much smaller difference in means than is observed with on-road test participants. Perhaps a different categorization of scores may benefit this measure.

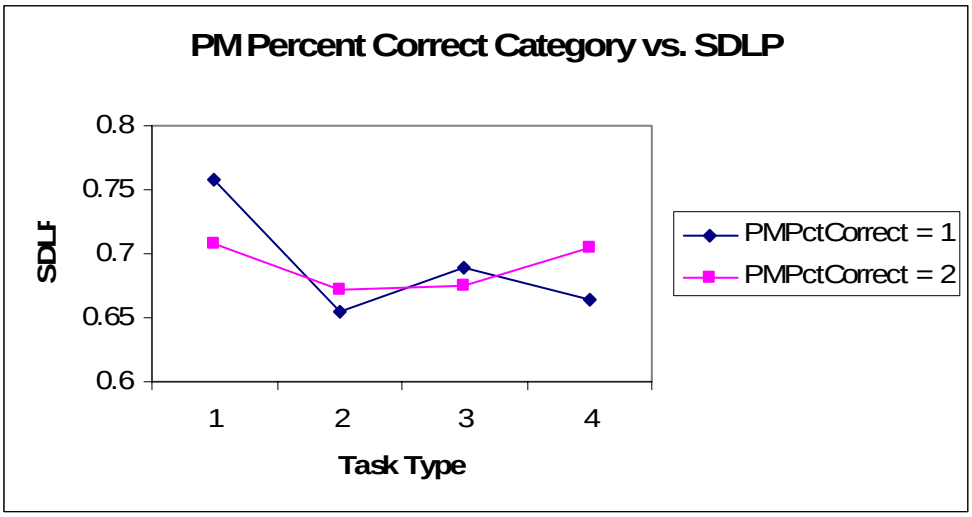


Figure 7-51. Patsys Percent Correct Versus Test Track Mean SDLP

7.5.4.2 PM Average Reaction Time

Patsys Manikin mean reaction time was categorized in three levels and plotted against on-road standard deviation of speed in Figure 7-52. There is little difference between the two groups with the fastest reaction times, and they follow the same between tasks pattern with the standard deviation roughly corresponding to task time. For the third group however, only Just Drive is similar to the other groups. For all other tasks, this group has higher standard deviations and the pattern between tasks does not follow task duration as with the others.

Figure 7-53 presents this effect for on road mean speed difference. Here all groups follow the same between tasks pattern, roughly following task duration. The effect shows with the slowest reaction time group having the highest speed difference for all but the Just Drive task types. The overall difference between groups is small but in the case of the third group, it is uniformly about two feet per second higher for the three task types.

This metric did not have a significant effect for the test-track data in part due to the distribution of scores, and in part, to the increase in variability of measures induced by the additional test-track only tasks.

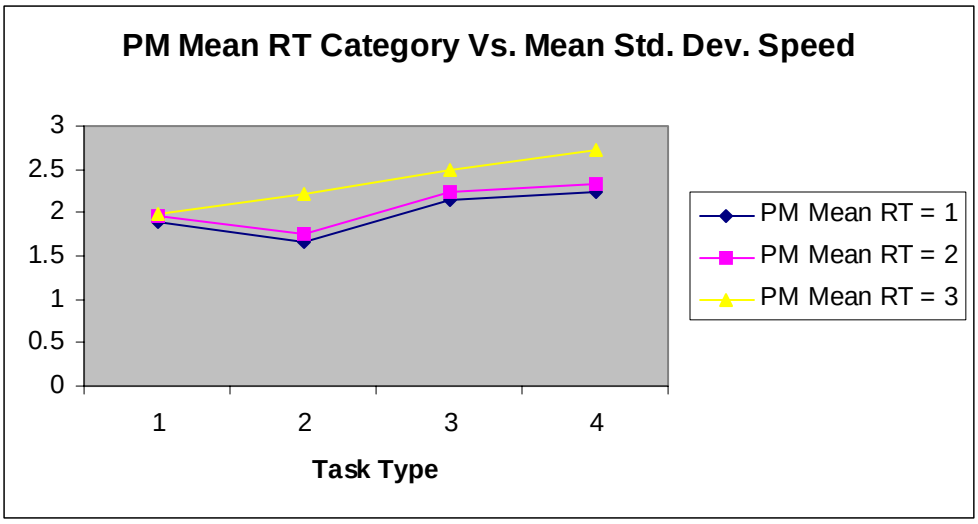


Figure 7-52. Patsys Manikin Mean RT Versus On-Road Mean Standard Deviation of Speed

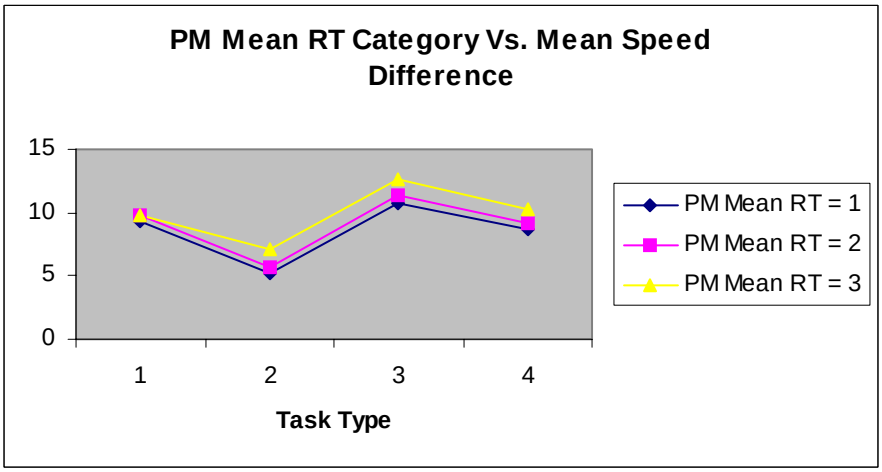


Figure 7-53. Patsys Manikin Versus On-Road Mean Speed Difference

Patsys Manikin percent correct has a significant effect for SDLP on the road but not for the test track. Not surprisingly, the lowest performing group for this measure also has the highest SDLP.

Patsys Manikin mean reaction time has significant effects with standard deviation of speed and speed difference for the road but again not on the test track. For both these measures, the group with longer reaction times had the most variable longitudinal control.

7.6 Summary of Age and Gender Effects and Cognitive Test Predictive Ability

Age has a significant effect on the performance of test participants in the areas of lateral and longitudinal fixed-base driving simulator control and on detection of various visual stimuli as well as processing speed for those stimuli.

Age also has a significant effect on lateral and longitudinal vehicle control. The decrement of this effect is most pronounced when older test participants are performing the tasks with a manual component. The youngest test participants are not necessarily the best performers, they are often nearly comparable to the middle age group, and in some tasks are comparable to the oldest age group.

Age has a similar effect for the test track venue as with the on road data. The additional tasks lead to more age-by-task interactions with the youngest test participants performing more poorly than the middle-age group for some task types.

Gender has an effect for multiple lab measures. STISIM lateral control indicates a gender-by-task interaction while longitudinal control shows female test participants have uniformly more variation across all tasks. Female test participants also have better detection and processing speeds for various visually presented stimuli.

Gender has a significant effect in lateral and longitudinal control measures. In both instances female test participants show more variability in vehicle control for all but the visual-manual task type where their performance is not appreciably different from that of the male test participants.

Results for the test track are very similar to those for on-road with the exception of male test participants having higher total lane exceedance counts for the visual-manual task type than did the female test participants.

Both Useful Field of View and Patsys tests differentiate test participants' performance based on their relative position in the distribution of scores. Differences in scores as well as means of vehicle statistics are relatively small. Some tests also exhibit task interactions, which may indicate a particular test is better suited to particular task types. Further investigation on how to apply these test scores individually to test participants rather than grouping classes of performers is desired.

7.7 Self-Rated Multitasking Ability and Ratings of Comfort and Confidence

Research conducted by Elander, West, and French (1993) points out that faster driving and tendency to commit traffic violations are associated with increased crash risk. These behaviors, in part, explain the relationships between age and crash risk and between gender and crash risk. Confidence in one's driving abilities and less concern about other's reactions to one's driving or concerns about risks while driving appear to be related to a more aggressive driving style. While the predictive power of these personality/social factors is not high, it does suggest that test participant assessment, in terms of confidence in one's own attentional capabilities, might be worthwhile.

A four-part, self-rated multitasking ability questionnaire was developed to obtain a subjective assessment of a participant's multitasking ability, frequency, and performance. See Appendix I for the complete multitasking questionnaire.

- The first part of the questionnaire gauged a participant's overall self-rated multitasking ability on a scale ranging from 1 to 7 with 4 as the mid-point. To assist the participant, 1 was rated as poor self multitasking ability, 7 was rated as outstanding self multitasking ability, and 4 was an average value.
- The second part of the questionnaire asked the participant to consider how often he/she tried to multitask, such as talking on the telephone, using the computer, or eating while driving. This rating scale also spanned from 1 to 7, with 1 as almost never multitask, 7 as very frequently multitask, and 4 as occasionally multitask.
- The third part of the questionnaire assessed the physiological state of the participant when he/she was engaged in multitasking. Four options were provided: frustrated, stimulated, anxious, and other. If the participant marked other, he/she was requested to specify his/her physiological state.
- The last part of the questionnaire related to the participant's performance when engaged in multitasking. The participant was requested to choose one of two available options regarding performance; his/her performance on both tasks is often as good as if he/she did them one at a time or his/her performance on one or the other task tends to suffer somewhat when he/she is doing two tasks together.

A comfort and confidence questionnaire was also developed in this study to evaluate the subjective rating of comfort and confidence of a participant in performing a task while driving. For each task, the participant was asked to evaluate four scenarios on a scale from 1 to 7 where 4 was the midpoint.

- The participant was first asked to assess his/her self-level of comfort in performing the task while driving. A rating of 1 indicated the participant felt completely uncomfortable, 7 indicated the participant felt completely comfortable, and 4 indicated the participant was neither comfortable nor uncomfortable.
- For the second question, the participant was asked to rate his/her level of confidence in the ability to perform a task while driving and maintaining speed, lane position, headway distance, and being attentive to changes in traffic scenarios and events on the road. A rating of 1 corresponded to the participant having no confidence in his/her ability to perform the task and drive safely, a rating of 7 meant the participant was very confident performing the task while driving safely, and a rating of 4 meant the participant was uncertain.
- On the third question, the participant was asked to rate the level of comfort he/she would experience if a driver in a near-by vehicle was to perform this task while driving safely. The rating scale for this question was identical to the one used in the first question, where a rating value of 1 was completely uncomfortable, 7 was completely comfortable, and 4 was neither comfortable nor uncomfortable.
- For the fourth question, the participant was asked to rate the level of confidence he/she would experience for a driver in a near-by vehicle performing this task. The participant had to assess the other driver's ability to perform the task while driving and maintaining speed, lane position, headway distance, and being attentive to changes in traffic scenarios and events on the road. The rating scale for this question was identical to the one used in the second question, where a rating of 1 corresponded to the participant having no confidence in the other driver's ability to perform the task while driving safely, a rating of 7 meant the

participant was very confident in the other driver's ability to perform the task while driving safely, and a rating of 4 meant the participant was uncertain of the other driver's ability to perform the task while driving safely.

7.7.1 Self-Rated Multitasking Ability Results

The results for the first question, Self-Rated Multitasking Ability, showed that there was no significant difference between samples of participants assigned to each venue (lab, on-road, and test track) in terms of their self-rated multitasking ability (see Figure 7-54).

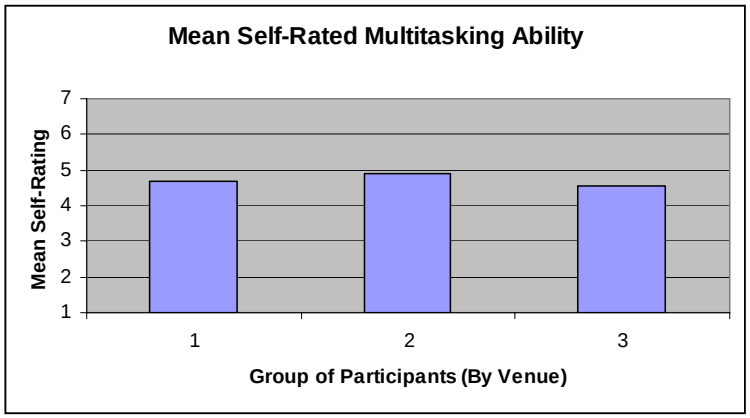


Figure 7-54. Ability to Multitask

1=Lab, 2=Road, 3=Track

7.7.2 Frequency of Multitasking Results

For Question 2, Frequency of Multitasking, there was also no significant difference between samples of participants assigned to each venue in terms of their self-reported frequency of multitasking (see Figure 7-55).

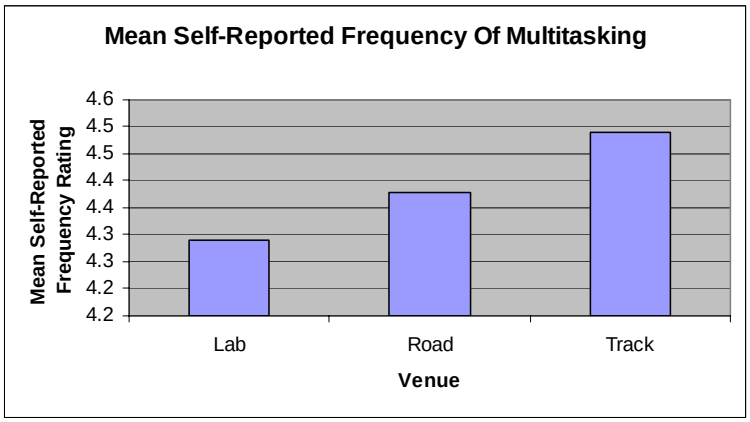


Figure 7-55. Frequency of Multitasking

1=Lab, 2=Road, 3=Track

The results showed self-ratings of multitasking ability did not differ significantly between age groups (see Figure 7-56).

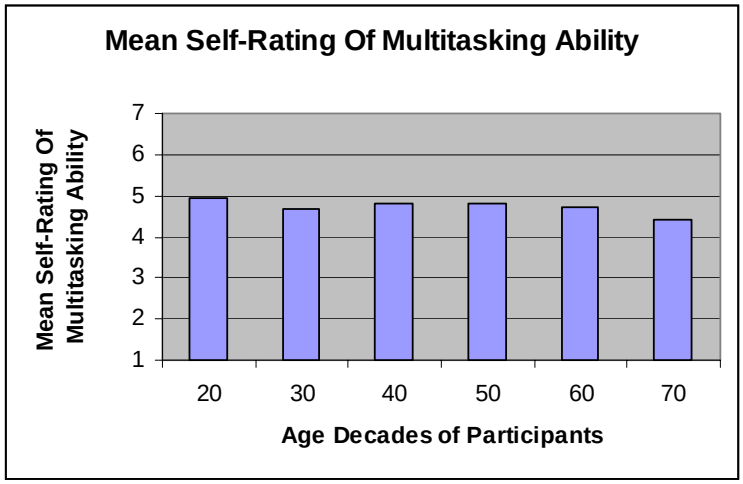


Figure 7-56. Self-Reported Multitasking Ability by Age

However, self-reported frequency of multitasking did vary significantly by age decade. Self-reported frequency of use by age (see Figure 7-57) showed that drivers in their 70s report significantly lower rates of multitasking than any other groups of drivers.

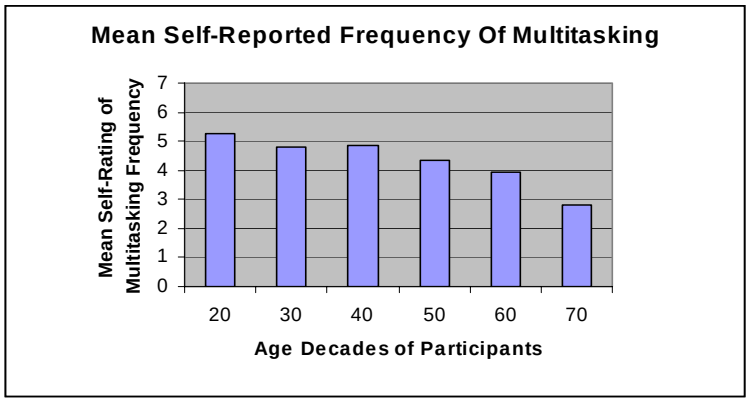


Figure 7-57. Self-Report Frequency of Multitasking

There was no significant gender difference in self-reported multitasking ability (see Figure 7-58).

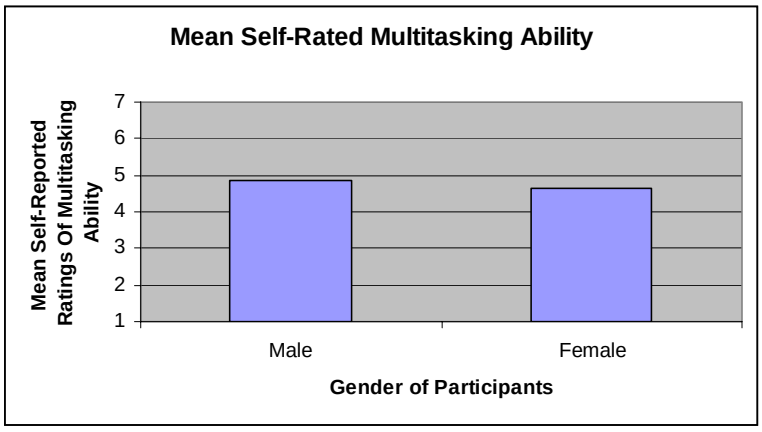


Figure 7-58. Gender Difference in Self-Reported Multitasking Ability

There was no significant gender difference in self-reported frequency of multitasking (see Figure 7-59).

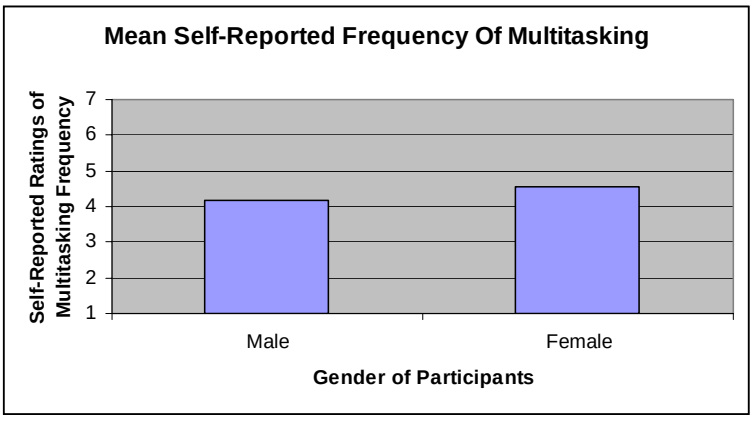


Figure 7-59. Gender Difference in Frequency of Multitasking

7.7.3 Self-Report of Feelings Results

Question 3 dealt with a Self-Report of Feelings when Multitasking: How Do You Feel When You Must Multitask? Participants were asked to rate themselves on a checklist of three adjectives (frustrated, stimulated, or anxious) plus "other" if those three terms did not capture their feelings. The other category was chosen more often than any other descriptor (see Figure 7-60).

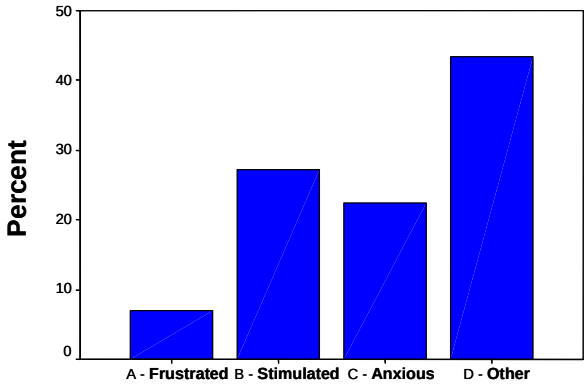


Figure 7-60. Self-report of Feelings

When the write-in comments for "other" category were further analyzed (in Figure 7-60), "calm" was the most frequent response (see Figure 7-61). Many of the participants felt "calm", even if this response was not available to them as a choice. Redesign of this question for future work may be necessary to include responses such as calm.

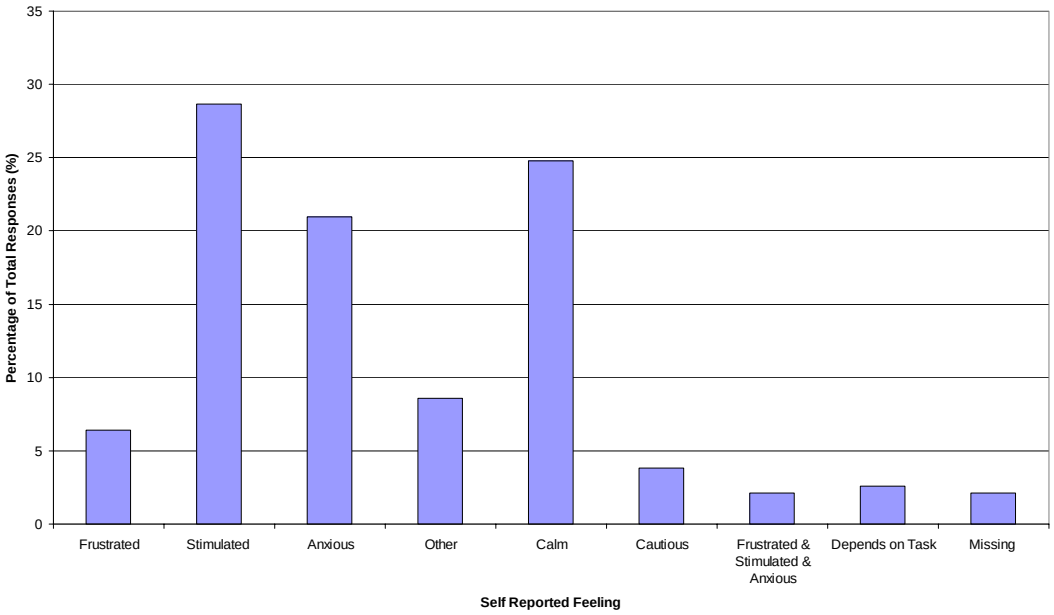


Figure 7-61. Self-Report of Feeling When Multitasking

7.7.4 Performance While Multitasking Results

Question 4 inquired about how participants' performance is affected by multitasking. Participants were asked: "When you must multitask, which tends to be most true of you?"

Participants could respond that their performance on both tasks is often as good as if each task was done one at a time or their performance on one or the other task tends to suffer somewhat when doing two tasks together (as compared to one at a time.)

More than half the participants felt that multitasking had no effect on performance (see Figure 7-62). However, the overall results of the DWM study showed some effects on performance of multitasking. It seems that most study participants underestimated the effect of multitasking on their actual driving performance.

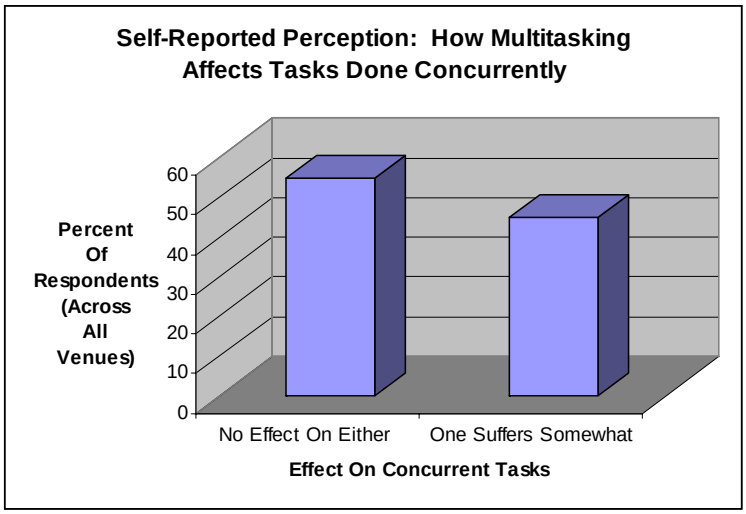


Figure 7-62. Effect on Concurrent Tasks

An analysis of the self-reported frequency of task engagement averaged across venue, age, and gender, found that conventional, visual-manual tasks such as HVAC adjust and radio tuning had a higher frequency of engagement than advanced technology, auditory-vocal and visual-manual tasks such as listening to flight information and navigation destination entry (see Figure 7-63).

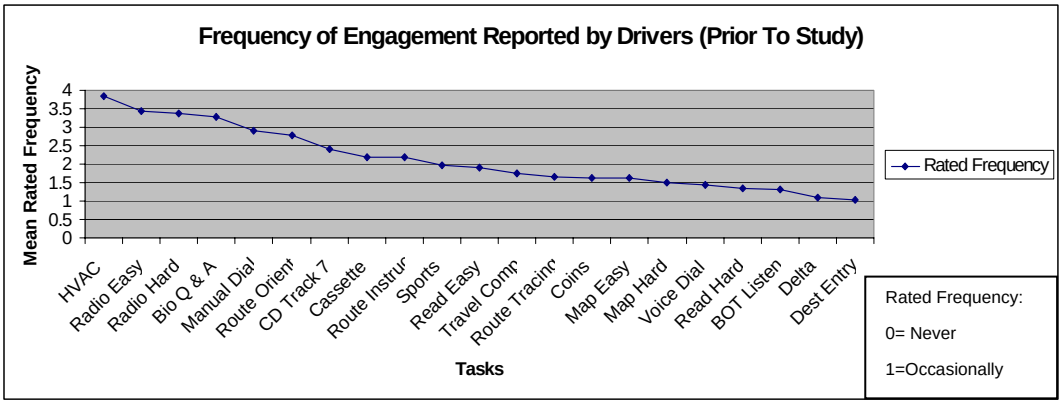


Figure 7-63. Self-Reported Frequency of Task Engagement

Age effects were present in Frequency of Engagement patterns for these tasks: HVAC, Radio Tuning, CD/Track 7, Manual Dial, Map (Hard), Route Instructions, Book-on-Tape Listen, Travel Computations, Read (Easy), Map (Easy), and Coins. Gender effects for the frequency of engagement patterns occurred in these tasks: Insert Cassette, Sports Broadcast, Route Tracing, Travel Computations, Read (Easy), Map (Easy), and Coins. There were interactions in the frequency of engagement patterns for the age by gender interaction for the Book-on-Tape Listen and Coins tasks, and the gender by venue interaction for the Route Orientation and Voice Dial tasks.

7.7.5 Comfort and Confidence Results

Question 5 requested a Self-Report of Comfort and Confidence while engaging in multitasking. This mini-questionnaire was administered during the intake portion of study (prior to task training and task testing). For on-road participants only, it was also administered as a post-test questionnaire.

This mini-questionnaire was intended as one source of information about participants' prior familiarity with tasks and possible willingness to engage in multitasking during subsequent parts of the protocol. For each task in study, the questionnaire asked participants to rate comfort or confidence on 7-point scale.

The first question for this mini-questionnaire dealt with comfort: How comfortable would you be with the ability of a driver in a car near you doing this task while continuing to drive safely? The participants were also asked to rate their own ability for this. The comfort scale tried to gauge "unease", and was intended to tap into the "feelings" aspect of the experiential side of multitasking. How uneasy the participant felt doing the task while driving would have been another way to interpret this.

The second question was with regard to confidence: How confident would you be with the ability of a driver in a car near you doing this task while continuing to drive safely? The confidence scale was intended to be more of a "thinking exercise" about how certain the test participant was of the other driver's ability to detect events, stay in the lane, control speed and headway, and do the subsidiary in-vehicle task. The participants were also asked to rate their own ability for this.

The original intent was for the comfort and confidence ratings to be thought of and treated separately by the participants. However, the scales for the comfort and confidence ratings did not end up being different. The data showed that respondents used the scales in the same way; the questionnaire did not achieve this intent to treat the comfort and confidence ratings separately. Due to the similarity of responses to the two scales, the data for the comfort and confidence ratings were combined.

For the pre-test questionnaire on comfort and confidence in multitasking while driving for a variety of subsidiary in-vehicle tasks, the participants consistently rated themselves more comfortable and confident in their own performance than other drivers (see Figure 7-64).

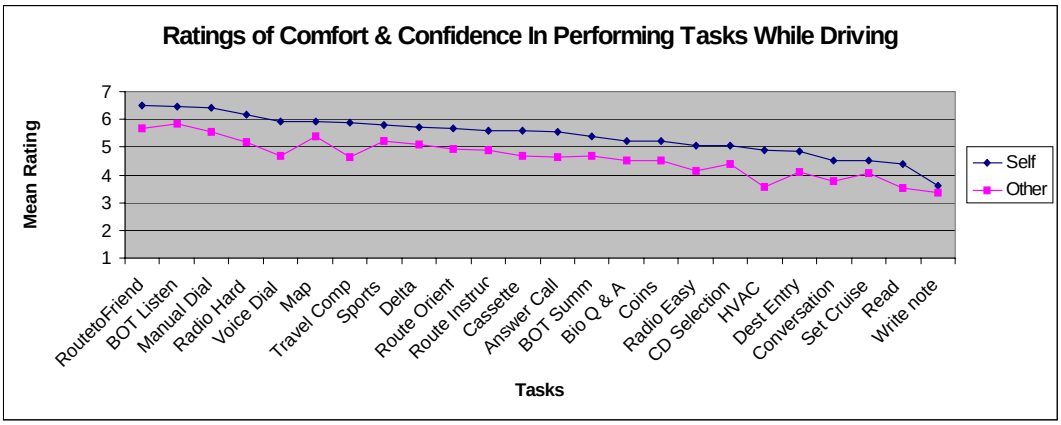


Figure 7-64. Ratings of Comfort and Confidence

Female drivers, in general, had higher comfort-confidence values than male drivers (see Figure 7-65). For some of the tasks such as Manual Dialing and Book-on-Tape Listen, the comfort-confidence values for the females were the same as those for males. But for some tasks such as a map task or navigation destination entry, the female comfort-confidence values were higher than males.

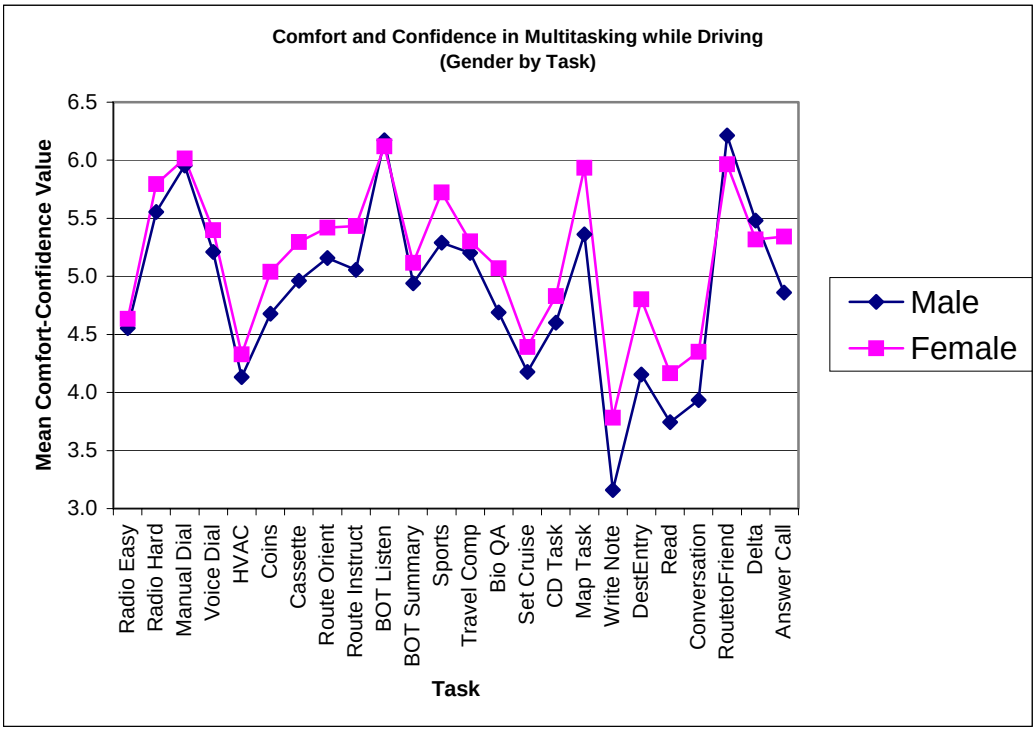


Figure 7-65. Comfort and Confidence by Gender

Older drivers had lower comfort-confidence values than younger drivers. Drivers in their 60s and 70s, in particular, responded with lower comfort-confidence values than drivers in their 20s and 30s (see Figure 7-66).

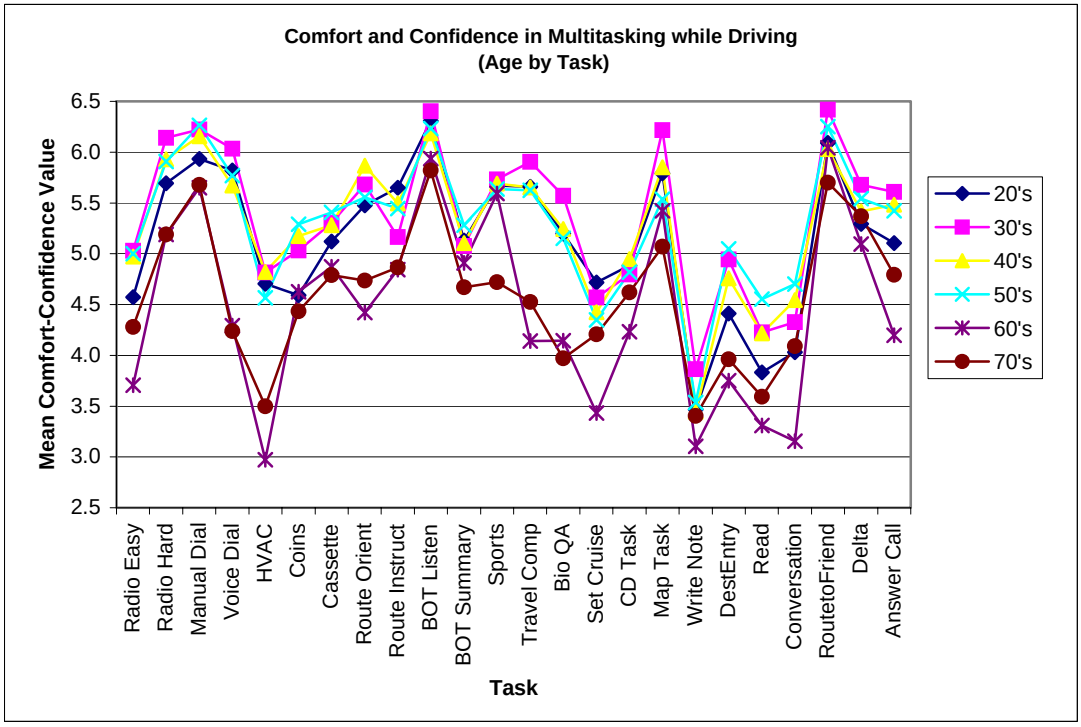


Figure 7-66. Comfort and Confidence by Age

When the ratings of comfort and confidence in multitasking while driving were compared in the pre- versus post-test conditions, the comfort and confidence varied for all subsidiary in-vehicle tasks (see Figure 7-67). Some tasks, such as Radio (Easy), produced higher ratings of comfort and confidence in multitasking while driving in the post-test condition compared to the pre-test condition. This suggests that this task was not as difficult as the panelist's prior expectations. Other tasks, such as manual phone dialing, showed the opposite effect: panelists were quite comfortable and confident in their prior expectations of completing this task, but lowered their comfort and confidence ratings after attempting the task.

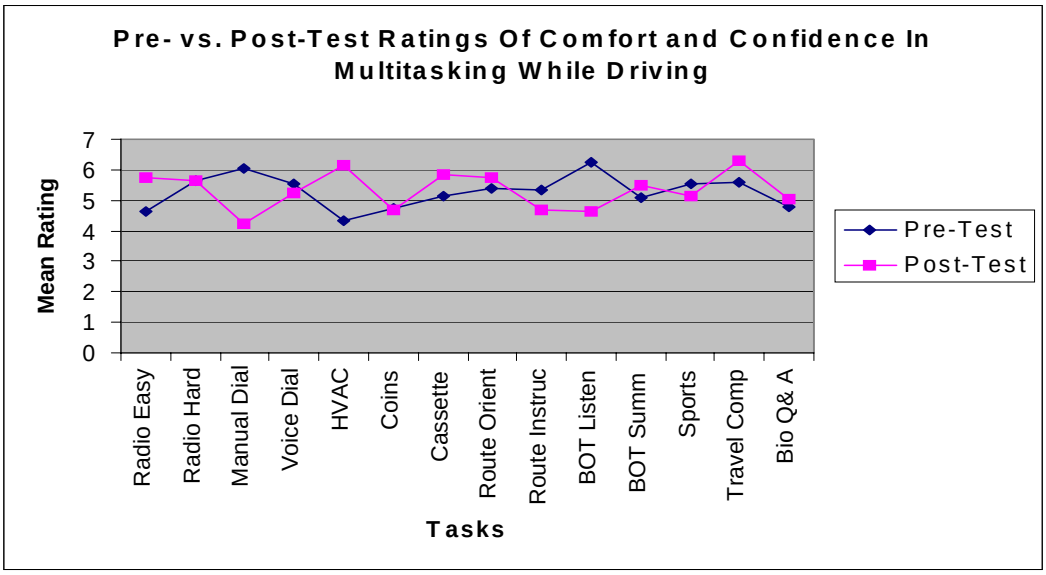


Figure 7-67. Pre- Versus Post-Test Ratings of Comfort and Confidence

7.8 Selection of Test Samples for Future Studies

As was shown in section 7.3, there are numerous age effects on various metrics of driving performance. There are also significant age-by-task interactions present in these driving performance metrics. For instance, the 60 to 79 year old age group typically has significantly more variation in lateral and longitudinal control than other age groups, especially for tasks that have manual components. The youngest age group, 20 to 39 year olds, often have the least variation in vehicle control, although this is not always true. For instance with SDLP, the youngest age group has variation nearly equal to the oldest participants for all but the visual-manual task type.

Section 7.4 detailed significant differences between male and female drivers for numerous driving performance metrics. Typically female participants had more variation in lateral and longitudinal vehicle control than male test participants. This was especially true for auditory-vocal and mixed-mode task types.

Section 7.6 showed a number of effects of age and gender for self-reported ratings. The oldest participant's typically reported less frequent use of in-vehicle devices and less comfort performing those tasks. The youngest test participants often reported the most use and most comfort in their ability to perform subsidiary in-vehicle tasks. While the oldest participants seemed to recognize the discomfort potential of in-vehicle tasks, the youngest test participants did not.

Given the performance differences between age groups and gender, as well as the self-ratings differences, test samples for product evaluations should be constructed to be balanced for all age groups as well as gender. Testing only a portion of the age range of potential users can misrepresent driving performance for the users who are most likely to use in-vehicle devices. For instance, a sample constructed of only the older half of participants would yield performance metrics that show more impact on the driving task, but would not accurately represent the oldest test participants.

Excluding the younger age groups from testing fails to represent the participants that are more likely to use in-vehicle devices and who are less likely to recognize the potential impact of these devices on their driving performance. While these younger participants often had the best vehicle control during tasks, several metrics showed that for some task types their performance was nearly equal to that of the oldest drivers.

7.9 Chapter References

Elander, J., West, R., and French, D. (1993). Behavioral correlates of individual differences in road-traffic crash risk. An examination of methods and findings. *Psychological Bulletin*, 113(2), 279-294.

8 Discussion and Recommended Toolkit

8.1 The Conceptual Context: Driver Workload and Distraction

In order to discuss the findings of the DWM project, it is important to place those findings within the conceptualization of driver workload and distraction that provided a framework for this research.

At the outset of the project, distraction was conceptualized as a mental and/or physiological state in which a human who was attending to one activity switches attention to another (without necessarily having conscious awareness of the switch and without necessarily having controlled or chosen the switch of attention). If the switch of attention can be made without degradation of the primary task, then there is assumed to be little to no distraction potential.

As applied in this project, the focus was on issues of distraction during driving that may be triggered by driver workload (with an emphasis on issues of overload and/or interfering workloads) that are associated with in-vehicle task activities. That is, of particular interest were states in which driver workload interfered with driving performance to a measurable extent. In that situation, it was hypothesized that this type of degraded driving performance might be interpreted to suggest that some level of distraction had occurred. Of course the state of distraction can be triggered by many things; and during driving, the largest single source has been identified to be outside the vehicle (Stutts, Reinfurt, Staplin, and Rodgman, 2001). Tasks can be associated with workload that leads to distraction. But not all distraction is associated with workload related to tasks. “Lost in thought” phenomena, awareness of stomach distress, seat discomfort, and road-side billboards are not workload but they may be distracting. However, another triggering source of distraction relates to driver workload from discretionary in-vehicle activities—overload, transitions from high- to low-workload, and “underload.” The focus for the project was on developing measurement systems that could detect driver workload that was on the overload end of the spectrum that interfered in observable ways with driving performance, and was associated with in-vehicle task activities. The project sought metrics that could detect the effects of task interference with driving performance reliably, with validity, and in a way that was practical enough for routine use within common product development processes.

The scientific measurement of mental and correlated physiological states is difficult in laboratory settings, let alone in dynamic functional environments of interest for traffic safety (in moving vehicles in real traffic). Distraction, as a transient and rapidly-changing mental state, was something for which no direct measure was available at the time the project was initiated; at least none that could be used in the dynamic environment of a moving vehicle on real roads.

Therefore, it was necessary to operationally define distraction in terms of observables that could be measured. The operational definition of distracted driving presented in Chapter 1, *Introduction* was that workload-related distraction (associated with in-vehicle activities) can be considered to exist when the workload associated with those in-vehicle activities leads to significant interference with or degradation of driving performance or eyeglance behavior.

Interference with driving was conceived to result from the competition for driver resources (psychomotor, perceptual, cognitive, attentional, and physical) between the driving task and concurrent subsidiary tasks, as manifested in degraded lanekeeping, longitudinal control, object-and-event detection (OED), or eyeglance behavior. The workload associated with concurrent activities was defined in this study to occur during the duration of a task. Drivers must allocate resources and attention dynamically across this time period to manage the execution of all tasks.

There is a level of workload associated with just driving the vehicle and detecting events, and there is additional competition for the driver's resources caused by the performance of an in-vehicle task and the associated need to manage the deployment of attention during multitasking.

Workload also has a facet called underload, i.e., insufficient activity to maintain normal performance. Underload can manifest itself in periods of "lost in thought" driving or other lack of focus on the driving task. This aspect of workload was outside the scope of this project.

The hypothesis was that degraded driving performance, degraded object and event detection, or changes in eyeglance behavior resulting from the effects of in-vehicle tasks would be discernible in such things as:

- changes in the variability of lane position or in the occurrence of lane exceedances;
- changes in speed or headway (mean or variance) during a task'
- changes in detection rates and detection times to OED events; and
- changes in the drivers' glance rates, glance durations, or percent of time glancing at certain locations in the visual scene.

Thus, in order to assess the effects of multitasking on driving performance, driving performance needed to be measured in four categories:

- lateral control (lanekeeping);
- longitudinal control (speed- and headway-keeping);
- object and event detection; and
- driver eyeglance behavior.

Some of the key initiatives on this project were to determine which metrics in each of the four categories were the best ones, that is, to identify metrics that were reliable and sensitive to discriminable degradation in performance introduced by task-related workload, since not all tasks increase workload. This is where univariate analyses played an especially important role. Chapters 3, 4, and 5 presented the results of the on-road, test track and laboratory phases of the study from the univariate perspective. As previously described, the objective of that analysis phase of the project was to identify which of the measures collected were repeatable, meaningful, and sensitive to changes in task-related workload.

Another key initiative was to discover configurations of interference across multiple metrics that were associated with distraction or, in the operational definition, with driving interference and degradation. This search for configurations of interference was approached from two different perspectives, a graphical approach and an analytical approach. Graphical multivariate methods (Chambers, Cleveland, Kleiner, and Tukey, 1978) were applied in the form of star plots. The star plots were used to provide a visual configuration of task effects on selected workload measures. These figures capitalize on human pattern perception to provide an at-a-glance summary of task effects in the form of stars whose arms represent different aspects of performance. Chapter 4 presented examples of star plots based on selected driving performance and eyeglance measures. This Discussion chapter presents examples of star plots derived from selected surrogate measures. Additional star plots are included in the Appendix S. Analytical multivariate analyses were to play an important role as well. In particular, Principal Components Analysis (PCA), presented in Appendix T, was attempted to uncover common factors or components behind multiple measures. However, due to the sparseness of the dataset on eyeglance and certain other measures, the outcome of a case-level PCA was unstable. Future work in this area will provide greater insights into the nature of the workload measures that enter into the analysis. The critical questions for

this endeavor had to do with whether states of driver workload, which produced overload or interference with driving performance, were manifest in just one of these categories of performance or were multiple categories of performance simultaneously affected? If multiple categories were simultaneously affected, were only specific patterns of effects characteristic of distraction? Were different patterns of interference/degradation across the categories of performance associated with, or diagnostic of, different types of distraction? That is, were there states of overload/interference that were specific to certain task types or driving conditions? The remaining sections of this chapter will discuss the findings that were observed from univariate analyses relative to questions of which metrics were the best ones for assessing task-related workload, as well as findings that were observed across multiple measures of driving performance and laboratory surrogate metrics relative to questions about driver workload itself and the nature of interference with driving performance.

The terms “degradation” and “interference” in driving have been used in several places throughout this report. These words beg corresponding questions. Degradation with respect to what? Interference relative to what? What is the basis of comparison? These are difficult questions that have not been satisfactorily answered.

Task effects as compared to just driving can be relevant and useful. This was the motivation and rationale for the Level 1 discriminability comparisons. Many studies of cell phone conversation while driving compare cell phone use to just drive conditions (Goodman, Barker, and Monk, 2005). However, some researchers have questioned just drive trials as a fair and equitable baseline for many applications (e.g., Dingus, McGehee, et al., 1995). Their reasoning is that just drive would often be too stringent a standard since many subsidiary tasks can be expected to impose at least some increased demand on the driver. A fairer and more equitable comparison for workload assessment may be to make comparisons relative to alternative implementations of the same function. For example, the workload effects of an electronic navigation system have been compared to those of a paper map or a list of text directions (e.g., Dingus, Hulse, et al., 1997). Other researchers have examined the effects of an in-vehicle system to effects associated with common in-vehicle activities that do not involve infotainment or telematics systems at all. For instance, Jenness, Lattanzio, O'Toole, and Taylor (2002) compared distraction effects associated with voice-activating dialing to the distraction effects of eating a hamburger while driving. The Alliance of Automobile Manufacturers (AAM) (2003) has taken yet another approach for comparative workload assessment. The AAM criterion selected was the operation of a single in-vehicle system (e.g., manual radio tuning), whose workload effects would be a benchmark against which to compare telematics device use.

A prior prediction approach was developed and implemented in the DWM research. Tasks were categorized as either higher-workload or lower-workload tasks based on prior research, theory, analytical modeling and engineering judgment. This approach used data that was not part of the DWM research in order to avoid circularity in interpretation. This method can be criticized on several grounds. There is one fundamental criticism, however, that should not be overlooked. The prior prediction approach provides a means to assess how consistent study findings are with earlier research. Prior prediction does not address the real-workload validity of those findings. At least an approximate “transfer function” is needed to relate workload measures or surrogates to actual distracted driving, not staged on-road or test track trials. Until that transfer function is estimated, there is no way to know what prior prediction assessment does. It may assess the degree to which current research findings are in accord with previous research. It is another thing altogether to assess which current and prior research findings are relevant.

8.2 Discussion of Findings from Univariate Analyses

8.2.1 Introduction

Many driver workload measures are described in the literature (e.g., Tijerina, Kiger, Rockwell, and Wierwille, 1996; Young, Regan, and Hammer, 2003). Selection among them might be driven by the following questions:

- How repeatable is the measure?
- How predictive is the surrogate metric of driving performance and eyeglance behavior across tasks?
- How discriminating is the measure to detect statistically significant differences between higher-workload tasks and lower-workload tasks that have been defined by prior predictions?
- How practical is the measure or surrogate to use?

This discussion reviews the development of the measures examined in the DWM project and the implications of the results obtained with them. It also discusses a number of interpretive issues and makes recommendations for the selection of surrogates.

8.2.2 Development of the DWM Object-and-Event Detection Methods

8.2.2.1 Pre-Pilot Testing on VIRTTEX

Ford Motor Company conducted a pre-pilot study of driver workload that related to the DWM project, although it was not formally a part of the project. The pre-pilot study was done on the Ford Virtual Test Track Experiment (VIRTTEX) moving-base driving simulator. It provided an opportunity to pilot-test an Object-and-Event Detection (OED) method for possible use in the DWM road study. The VIRTTEX method used events both in front and rear of the subject vehicle during a car following scenario. The front detection event was one in which a vehicle ahead of, and largely hidden by, the lead vehicle would sometimes swerve suddenly onto the right shoulder or the left lane and return immediately after straddling the lane line half a car width. The rear detection event was one in which a vehicle following behind the subject vehicle would suddenly swerve in a manner identical to that of the vehicle ahead of the lead vehicle. The participant's task was to use the turn signal to indicate detection and direction of a lane violation. The method in general appeared promising and the results of that study have already been published (Greenberg, et al., 2003). More recently, Reyes and Lee (2004) developed a variant of this method. In their study, a lead vehicle braked randomly in one condition. In another condition, the lead vehicle braked in response to a vehicle ahead of the lead vehicle that suddenly changed lanes into the lead vehicle travel lane. They found the latter method to be sensitive to distraction effects that were missed by techniques that involved lead vehicle deceleration alone. The Reyes and Lee study was also done in a driving simulator.

A platoon concept, derived from the VIRTTEX study, was retained in the DWM road and track trials. The platoon consisted of a lead vehicle, the subject vehicle, and a follow vehicle. It was logistically infeasible to have a fourth vehicle ahead of the vehicle ahead. Thus, the lead vehicle and the follow vehicle were to perform the swerve maneuvers to the berm lane. Pilot testing on the road revealed the swerve procedure to be infeasible. Safety protocols for on-road driving trials were put in place to conduct the swerve maneuvers with less lateral extent and only when surrounding traffic was sufficiently separated from the test vehicles. The test vehicles were clearly marked as such by placards. Only swerves to the berm lane were performed. The Michigan State Police was contacted to solicit their cooperation as well. Despite these precautions, surrounding traffic was still disrupted (as surmised by honking horns and other

indicators). This OED procedure was dropped, and Lead Vehicle Deceleration (LVD), Center High-Mounted Stoplight (CHMSL) onset, and Follow Vehicle Turn Signal (FVTS) onset events were developed as OED events instead. The platoon concept was retained. The importance of a follow-vehicle OED event was retained.

8.2.2.2 Further Analysis of the Lead Vehicle Deceleration Event

The Lead Vehicle Deceleration (LVD) event was an appealing stimulus for detection during car following. It was appealing because it had apparent relevance to rear-end crash occurrence. Nonetheless, analysis indicated that it may be incompatible for use with shorter tasks. Figure 8-1 presents box plots of the maximum optical looming levels for detect versus no-detect trials by task on the test track. Mortimer (1990, 1993) has reported that the threshold for optical looming is approximately 0.003 radians per second. The vertical line in the figure indicates this threshold. Overall, the figure indicates that shorter tasks are more likely to have had no-detect trials that were below the looming threshold than longer tasks. LVD had not built up to above-threshold levels during the period of short-duration task performance.

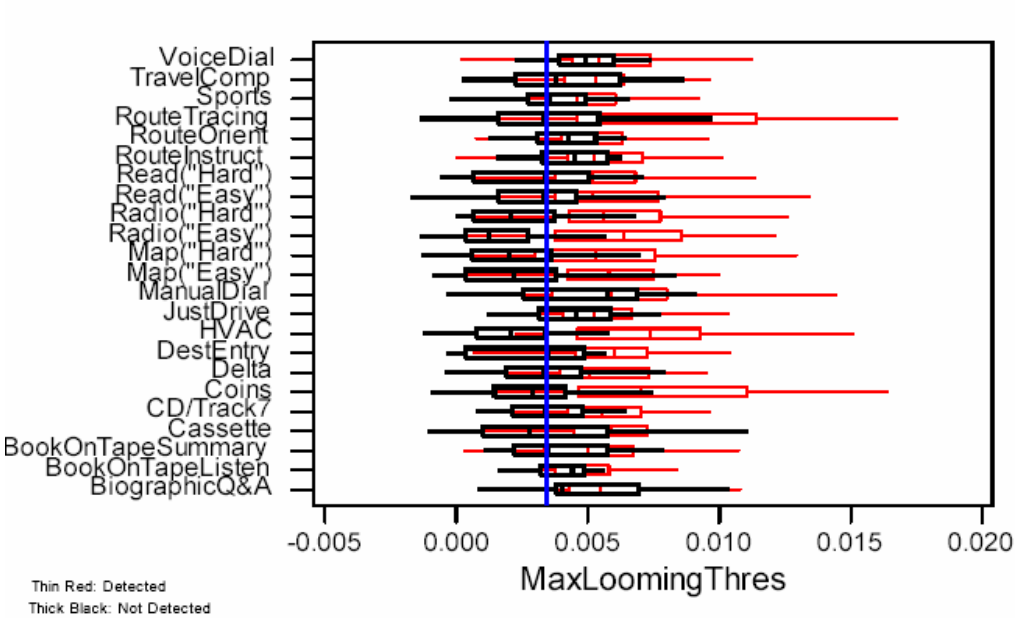


Figure 8-1. Optical Expansion Rate for Detected Versus Not-Detected Trials by Task for Track Trials

This explanation leaves unanswered the question of how there could be any detection below the looming threshold. Thresholds are defined at the 50th percentile of a sample of observers. Thus, a plausible hypothesis is that some of the test participants were more sensitive to looming cues. Another plausible hypothesis is that other cues were used at times. Mortimer (1992) pointed out that there are other visual cues to closing besides looming. Other cues include change (rather than rate-of-change) in headway distance or visual angle or area of the object ahead. The idea behind these other cues is based on the Weber ratio. The Weber ratio is the ratio of one stimulus magnitude to another magnitude that is just noticeable. If the difference in separation from one perceptual moment to the next is great enough, the driver can detect closing even if the looming cue is below threshold. The Weber ratio for headway separation is between 0.09 and 0.17, depending on various factors like driver individual differences in sensitivity, the type of visual cue, ambient light conditions, and so forth. So, if 100 feet separate the vehicles in the first perceptual moment, the separation would have to be 88 feet or less in the next perceptual moment

for the driver to detect the change in separation if the driver's Weber ratio was 0.12. It is difficult to apply in the Weber ratio concept in this project, even though it is empirically established. It is not easy to determine when the last “perceptual moment” and next “perceptual moment” were in time, especially if the perceptual moment is an internal cognitive sampling phenomenon (see Card, Moran, and Newell, 1983 for a discussion of this in a cognitive framework). It may sometimes be simply the change in the driving scene from one road eyeglance to the next. Regardless, the LVD event was complex to stage so as to present an above-threshold LVD event for detection during shorter tasks. For this reason, mild decelerations are not recommended for use with shorter duration tasks. A planned LVD event's duration to reach above-threshold looming conditions can provide an analytical means to determine if LVD events are suitable for a given task.

8.2.2.3 Task Duration and Response Window Issues for OED

OED responses had to be made before the end of the task to be counted as detections. This was done to minimize the chance of such measures being affected by performance shaping factors unrelated to the in-vehicle task itself. Examples of such extraneous factors between trials were activities like talking to the experimenter, adjusting one's seat position, listening to experimenter instructions, catching up to be in position for the next task, and so on. Yet other extraneous performance shaping factors include an increasingly strong stimulus (e.g., looming) or another stimulus that persists after the task is completed. Regardless, the response window was smaller for shorter tasks than for longer tasks.

Paradoxical results emerged for OED performance. Shorter and ostensibly less distracting tasks were associated with poorer detection performance than longer and ostensibly more distracting tasks. This might have been an artifact of the response window. Extraneous factors aside, there might have been detection during the task but the response was made afterward, perhaps because there was no sense of urgency to make the response.

To test this hypothesis, a separate analysis of detection performance was made to look for responses with an additional five seconds beyond the task end marker added to the response window. Only the CHMSL and FVTS trials were examined since the LVD provided a non-constant stimulus. The additional detections were primarily, though not exclusively, for shorter tasks. However, test track results (see Chapter 3) analyzed in this way revealed additional detections but did not fundamentally change the paradoxical pattern of OED results. The issue about how to address the OED task duration paradox is considered in the next section.

8.2.2.4 Possible Methods to Address Task Duration Effects on OED Results

OED results revealed that shorter (and presumably less demanding) tasks were associated with poorer performance than longer tasks. This is similar to a finding reported by Young and Angell (2003). In a multivariate analysis of performance measures taken from visual manual tasks performed while driving, they found a dimension of performance that was associated with attentiveness to events that was orthogonal to a dimension associated with task workload. This was associated with tasks which were low in workload, as characterized by few glances off the road, few lane deviations, low subjective workload, high subjective situation awareness, yet paradoxically high miss rates and long reaction times to outside events. In the DWM study, this result is found both for OED methods that present multiple stimuli per trial (laboratory OED methods) and OED methods that present only a single stimulus per trial (e.g., track methods). This result holds for simple light stimuli presented either alone or while driving in a simulator such as a peripheral detection task (PDT) Alone, or PDT performed with a simulator. It holds for road signs to be compared with a memorized set (Sternberg Spatial and visual stimuli). And it holds for CHMSL and FVTS onsets from other vehicles while driving (road and track methods).

This trend is not affected by the response window if an additional five seconds is added for the road and track trials.

The shorter, variable-duration DWM tasks were predominantly visual-manual tasks. However, the auditory-vocal Book-on-Tape Summarize task and the Voice Dial task were shorter tasks, and they too typically showed poorer OED performance than longer auditory-vocal tasks. It is possible that shorter tasks are as much or more resource-intensive than longer tasks. Detailed eyeglance data analysis (e.g., analysis of time series) may provide a mechanism to explain such effects. Such a mechanism would presumably go beyond the simplistic observation that longer tasks were associated with longer total time, in seconds, spent looking at the road scene.

One possible contributing factor to the paradoxical OED results is an experimental artifact of scheduling OED events regardless of task duration. The DWM trials were scheduled to present roughly the same OED stimuli for each task trial, short, long, and in-between. Therefore, OED results must be carefully interpreted because they do not explicitly factor in a task's duration. That is, the OED results make all tasks look alike in duration even as they appear different in detection performance. This raises an interpretive concern. Even if a shorter task may be resource-intensive, it is over sooner and this needs to be recognized along with other aspects such as the level of workload. Otherwise, the OED results may be interpreted to mean that tasks should be designed to last longer. Generally, this seems like inappropriate design, provided shorter task designs do not promote excessively longer task-related glance durations (e.g., two seconds or longer) or create usability problems.

Several methods might be used to address the task duration issue. One is to decide that tasks below a given duration should be given a pass on this variable. This use of task duration, with a threshold, is a key element of SAE J2364. The controversy over such an approach has been sufficient to be documented (Foley, 2005).

Another possibility is to match a task against a comparison task of a similar duration. For example, one might compare OED results for a 10-second task against the OED results for 10-second Just Drive trials during which OED probes were also presented. One problem with this approach is that any subsidiary task activity might be expected to impose more load than an equivalent period of just driving. The meaningfulness of task trial comparisons to Just Drive trials is therefore questionable. Another problem is how to compare task results to other tasks. Just Drive might be conveniently cut up into arbitrary lengths. It is not so easy to do so for subsidiary tasks. For example, what 10 seconds slice from a 90-second Destination Entry task should be compared to the demand of a 10-second HVAC task? Making long tasks shorter is conceptually appealing. However, that will often involve interrupting (terminating) a long task to compare it with a short (completed) task. This procedure would unfairly compare completed tasks with unfinished task segments. It could also complicate the test protocol considerably as one considers how many ways there are to segment a long task. Should it be at the beginning, near the middle, or toward the end of a longer task?

A third possibility is to rely on expert judgment to assess the distraction potential of short-duration tasks. Such judgments might be based on task or equipment characteristics. What these characteristics are and how they are best evaluated might be addressed by checklists prepared for that purpose (Stevens, Board, Allen, and Quimby, 1999). It is unclear how precise existing checklists might be to predict OED impacts. Checklists might also be augmented by measurements or estimates of single-glance durations that are often unusually long, e.g., greater than two seconds. However, glance duration data currently hard to obtain and may thereby not be considered practical.

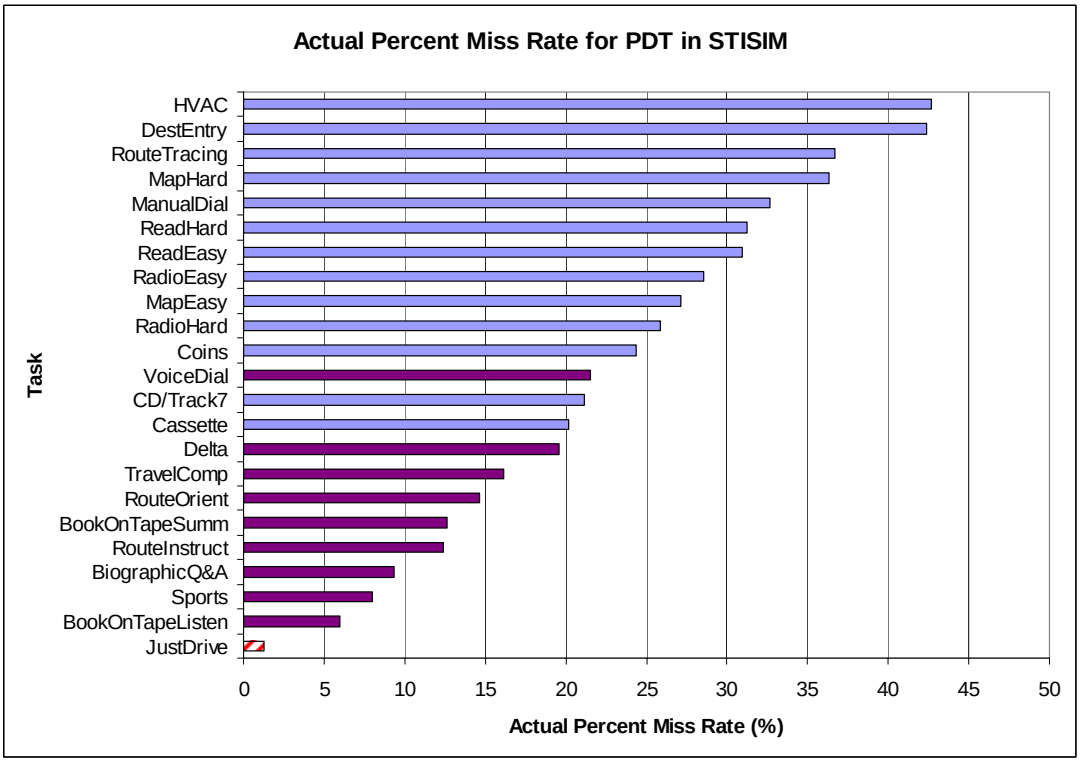
Yet other approaches might provide interpretable OED results over a range of task durations. One method is to ask the test participant to repeat the short task over and over for a longer duration.

For example, Mattes (2003) proposed a Lane Change Task or LCT, which has been called into question because it involves performance that makes a short task artificially long. A visually intensive task done repeatedly may artificially elevate a short task's workload measures (Noy and Lemione, 2001).

One might take the results at face value and assume that longer tasks are done at a more casual pace. This seems incomplete because longer tasks had more components in the DWM research. Indeed, that was why they were longer. Another interpretation is that the findings result from people being rushed by the experimental procedure to complete the requested task. However, since the participant instructions were the same for all tasks, this is unlikely to account for the durational differences.

One particular method that will be illustrated here is to make a task duration adjustment. The idea is to adjust the OED Miss Rates for different tasks by their duration. One method that will do this is to define a “duration adjustment multiplier.” This multiplier can be defined as the ratio of a task's typical duration to the typical duration of the longest task or a baseline task. This ratio will be a number between 0.0 and 1.0 that, when multiplied with the Miss Rate, provides an adjusted Miss Rate. If the range of typical task durations is small, the range of multiplier values will be small. If the range of durations is large, the range of adjustment factors will be large. One or more comparison tasks might be included in each evaluation or study for consistency of results. Or a constant value of “maximum duration” might also be applied. In any event, the multiplier is applied to the aggregate detection performance associated with a task. This is a coarse method that does not address momentary fluctuations in task demand over the course of its completion.

An example application of a duration adjustment multiplier to PDT-in-STISIM Miss Rates is provided in Figure 8-2. The figure shows both the unadjusted values and the adjusted values per task. The adjusted values more closely match prior predictions based on the literature and modeling efforts. The range of adjusted values for shorter tasks is compressed because the longest typical task duration value is so much larger. The generality and robustness of such an adjustment is unknown but merits further consideration.



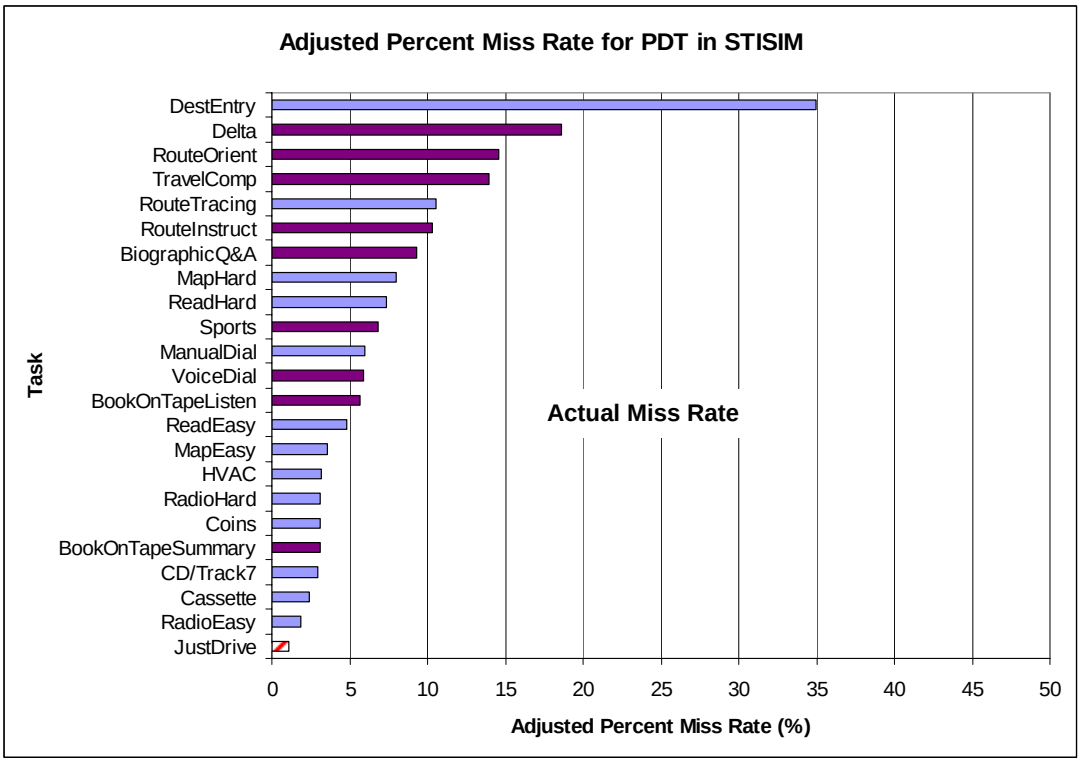


Figure 8-2. Effects of Duration Adjustments on PDTS Miss Rate Results

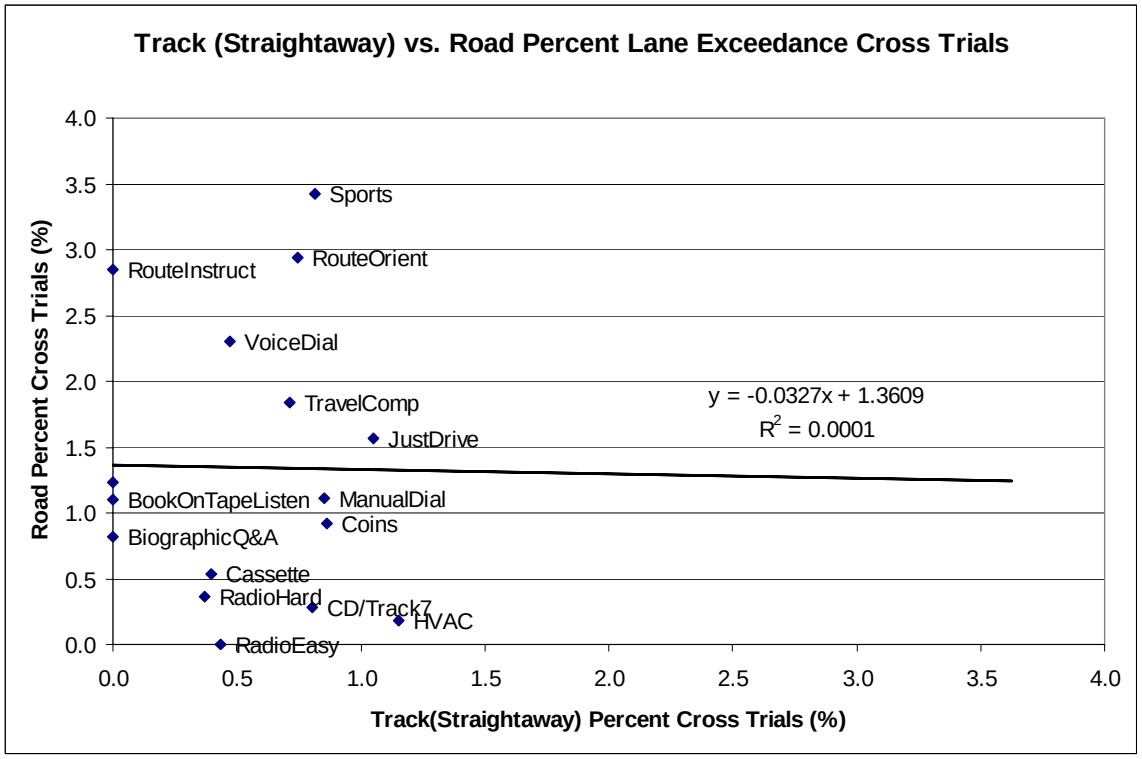
OED is a critical part of driver distraction assessment. Methods are needed to assess OED performance over a broad range of tasks. The DWM project has explored new ground with its various OED methods. These methods, properly interpreted, can provide empirical support for system design and implementation decisions. Several methods to account for different task durations in OED results have been offered. This will be an important area for further research. A traditional approach in workload assessment is to use task time itself as the workload indicator. Shorter duration is generally better. This is the rationale behind computational cognitive models (Card, Moran, and Newell, 1983) and time-and-motion analyses (e.g., Niebel, 1976) applied to task design. In the interim, the duration adjustment procedure may be a viable option for workload estimation until other, superior, methods are developed.

8.2.2.5 Differences between Road and Track Results – Effects of Curve Negotiation

Over all tasks, comparisons of road results versus track results indicate good agreement for the tasks and measures common in both venues. These tasks were limited, however, to the (predicted) less demanding visual-manual tasks, the auditory-vocal and Just Drive tasks, and the Voice Dial mixed-mode tasks. The major exception to this generalization was in the lane exceedance results.

The road and track lane exceedance differences might have been due to track geometry. The test track was an oval track. It had curves that gradually increased in lane width to 13 feet from the 12 feet width in the straightaway sections. Curve negotiation would also have required different steering than the straightaway sections. This hypothesis was tested by culling lane exceedances on curves from the track data and then re-computing the task-level summary statistics (Percent Lane Exceedance Trials).

Figure 8-3 presents the lane exceedance results between Road and Track lane exceedance results both for all track data (upper figure) and for track data after the curve data are removed. The differences remain. Lane exceedances nevertheless can be important indicators of workload and should not be ignored when they occur. They are practically appealing because the exceedances can be counted by an observer without sophisticated equipment (e.g., in real time). Measures such as standard deviation of lane position (SDLP), peak lane exceedance, and integrated area out of the lane require a robust lane tracker and suitably marked roads.



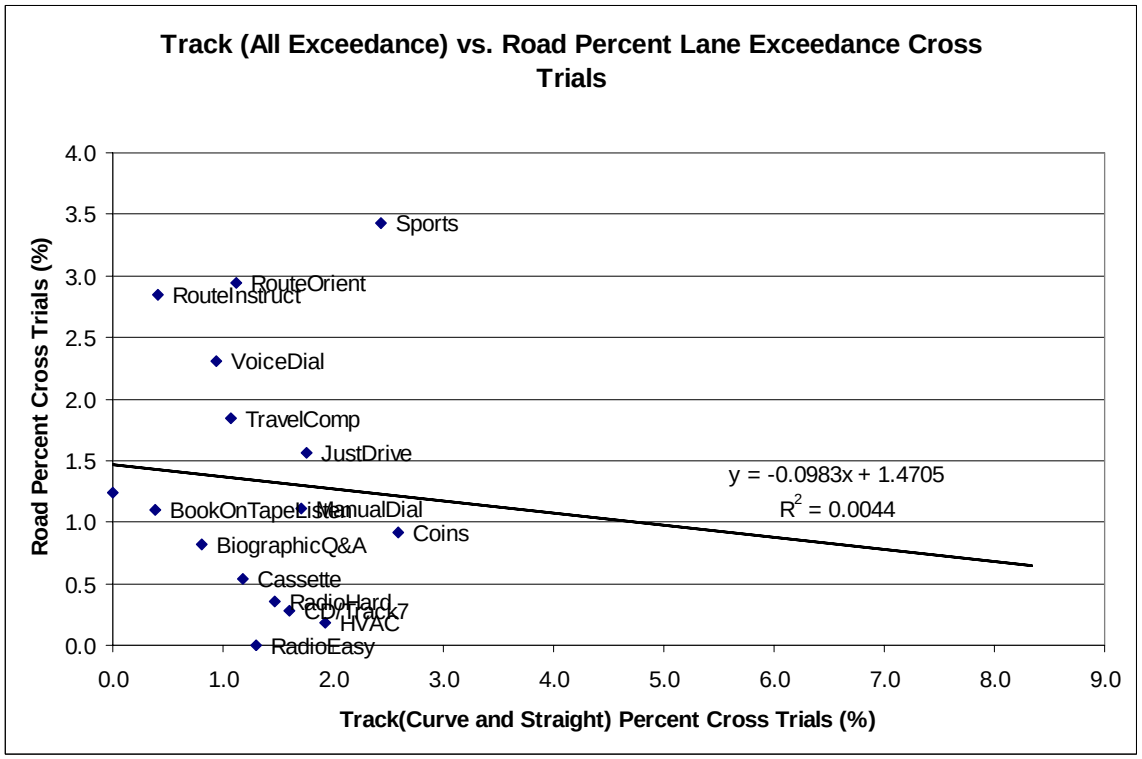


Figure 8-3. Road Versus Track Lane Exceedance Results for All Track Data (upper figure) and With Curve Data Removed (lower figure)

8.2.3 Selection of Driver Workload Measures

8.2.3.1 Issues Related With the Selection of Laboratory Surrogates

The results in Chapter 5, *Laboratory Results* reveal several interesting and important points about estimating the distraction potential of subsidiary tasks while driving. The most obvious point is that there is no single, perfectly discriminating measure of driver workload applicable to all tasks. But some workload measures are better than others. It appears to depend on the type of task. For example, auditory-vocal tasks can be distinguished with OED surrogate measures. Lateral and longitudinal control measures and traditional eyeglance measures are not particularly useful to discriminate between lower-workload and higher-workload auditory-vocal and Just Drive tasks.

Visual-manual tasks have a different profile of preferred measures. Because of the paradoxical effects of task duration on OED performance, OED measures are sometimes hard to interpret, especially for shorter tasks. Various options to deal with this task duration paradox have been presented but none seems foolproof. On the other hand, visual-manual tasks can be well discriminated with selected lateral and longitudinal control and eyeglance measures, whether obtained in actual driving or through laboratory surrogates. As indicated in other parts of this report, these are duration-driven measures best suited to tasks with task-intrinsic durations. These are tasks for which their duration ends when the goal is reached rather than after an arbitrary length of time.

Another important aspect of the surrogates is their discriminability as compared to road or track measures. The laboratory surrogates generated larger or more discriminable effects than were found on road or track. While predictions between surrogates and driving performance measures were often directionally correct, the in-vehicle effects were sometimes not big enough to stand above an otherwise noisy measurement process. In any event, laboratory results did not necessarily yield the same outcomes as results obtained in actual driving. This is an important point to bear in mind. Studies conducted on simulators should be validated or compared with real driving trials.

The last sentence above introduces a quandary of its own. Road and track trials in studies like the DWM study trials are simulations too. Test participants volunteered to come and work with strangers. They were asked to learn sometimes unfamiliar tasks. Then they were asked to perform these tasks while driving a designated route not of their choosing. The test vehicle was highly instrumented, unfamiliar, and staffed by strangers. Lead and follow vehicles, also from the study, drove along at about a requested speed of 55 mph. Some test participants followed the lead vehicle on a highway with a 70 mph posted speed limit. Other test participants followed the lead vehicle on a high speed oval test track at a proving ground. It is possible or even probable that test participants were not driving like they normally would. It is a guess, but participants under such conditions might have been highly vigilant and tried very hard not to make mistakes or appear foolish. The participant's perception of risk behind the wheel of a real car under such circumstances might be high.

Artificiality was also present in the laboratory, along with social motivation to appear sharp and conscientious. Yet laboratory surrogates sometimes showed task effects that road and track trials did not. It is not clear why this should be so. Perhaps laboratory participants did not care so much about their performance in laboratory trials as did participants in road or track trials. Clearly, there was no risk of property damage or bodily harm if a participant missed a Sternberg probe now and again or departed the lane in the simulator. Perhaps the laboratory trials themselves imposed more load than actual driving. For example, driving a fixed-base simulator like STISIM can impose more visual demand because there are no motion cues to signal lane drift or deceleration to the driver. Lab participants perhaps were not motivated to get through an

occlusion trial quickly while performing a given task. Participants who performed that same task while actually driving may have been more motivated to get the same task done quickly.

The quandary comes down to two very different interpretations of these results. On the one hand, laboratory surrogates may have generated false positives or false alarms because they were not “real-world”. On the other hand, road and track trials may have resulted in false negatives or missed detections because of a perception of risk unlike what would be present in natural driving. The DWM data do not provide a means to answer this question.

One final comment on the interpretation of surrogate measures is related to the transfer function issue. Even with statistically significant results in hand, interpretive issues remain. Normally, a statistically significant result can be intelligently followed up with the question "But is it practically significant?" In the case of DWM task-related driver distraction assessments, it is not possible to answer this question one way or the other.

Improved distraction assessment will require more insight into exposure factors that help better determine the true impact of a task on highway safety. A study that examines frequency of use, location of use, time-of-day of use, etc. will complement the workload demand measures evaluated in the DWM project and lead to a better assessment of the true nature of the distraction problem in driving.

8.2.4 A Multivariate Graphical View of Laboratory Surrogates

Multivariate graphical methods have proven very useful in exploratory data analysis. One such method is the star plot. A star plot with *p* arms is a multivariate view of an object along *p* measurements. An example of how a star plot might be applied to the laboratory surrogate measures is presented here.

The first step in preparing star plots was the selection and placement of the measures of interest. The example prepared for this section used five dimensions. These five dimensions were meant to represent aspects of real-world driving to which they might be related. These measures were the following:

- STISIM Task Duration (as a surrogate for task duration while driving)
- STISIM SDLP (as a surrogate for continuous lanekeeping)
- STISIM Percent Lane Exceedance (Cross) trials (as a surrogate for discrete lapses in lanekeeping)
- Speed Difference (as a surrogate for longitudinal control)
- Total Shutter Open Time (TSOT) (as a surrogate for task-related total eyes off road time and glance counts)
- PDT-in-STISIM (PDTS) Percent Miss Rate (as a surrogate for OED while driving).

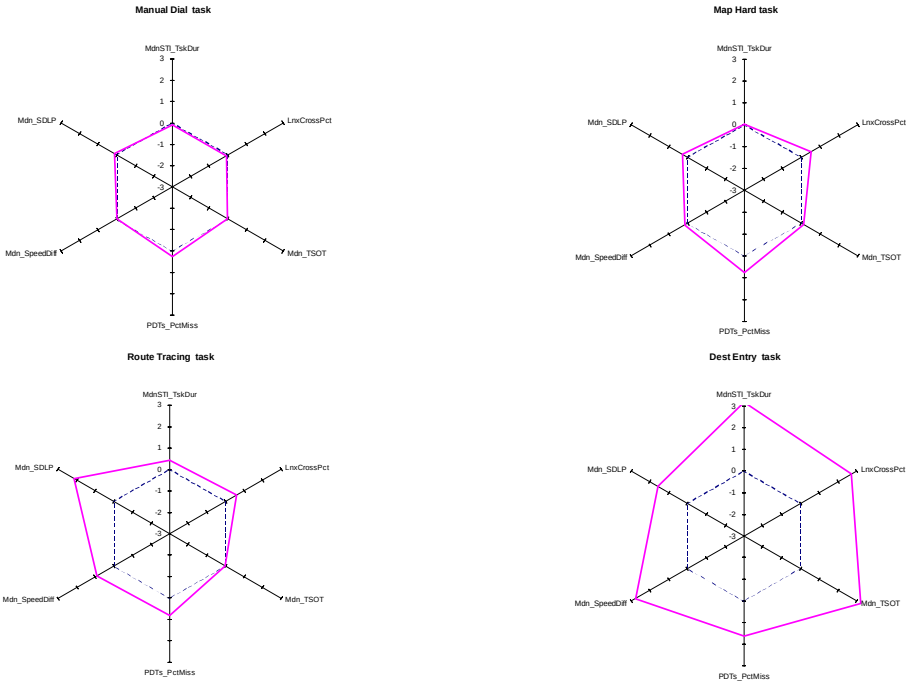
The next step was to standardize these different measures into a common unit for plotting and analysis. Otherwise, it would be difficult to comprehend the length of one arm of the star given in feet, another given in seconds, another given in percent, and so on. Therefore, standard scores were calculated, using task-level summary statistics from all of the visual-manual tasks. For a given measure, this involved calculating the mean and standard deviation for the set of tasks, then subtracting the mean from each task's score and dividing that difference by the standard deviation. For example, Median STISIM Task Duration was the measure. The mean and standard deviation of the 13 visual-manual tasks' median STISIM task durations were calculated. Then that

mean was subtracted from, say the median STISIM task duration of HVAC. The HVAC difference was then divided by the standard deviation of the visual-manual task set. This method was applied for each measure. The result was, per measure, a transform into dimensionless units with a mean of zero, a standard deviation of 1, and no change in the shape of the original distribution.

The last step was to assemble the star plots. This was done in Microsoft Excel using the Chart Wizard. Each star plot was set to have annuli or rings at: –3, -2, -1, 0, 1, 2, and 3, which correspond to plus-or-minus three sigma. The zero ring or annulus was highlighted for reference as a dotted line. The zero ring represents the mean across all tasks in this set but should not be taken to represent a criterion or norm of any type. Then the actual value of each standardized measure for a given task was plotted in a thicker, solid line in a different color.

Examples of star plots for selected visual-manual tasks are provided in Figure 8-4 and Figure 8-5 for higher-workload and lower-workload tasks, respectively. Consider the higher-workload star plots first. The Manual Dial task is unique because it sits on the borderline demarcated by the mean annulus on all dimensions. The Map (Hard) task appears to have values above the norm for PDTS Percent Miss Rate and the lanekeeping measures, but not for task duration, TSOT or Speed Difference. By contrast, the Route Tracing task has above-normal lanekeeping measures (SDLP and Lane Exceedance), poorer PDTS performance, and longer task time. But its TSOT value is not above normal. The Destination Entry task is high on all workload surrogates.

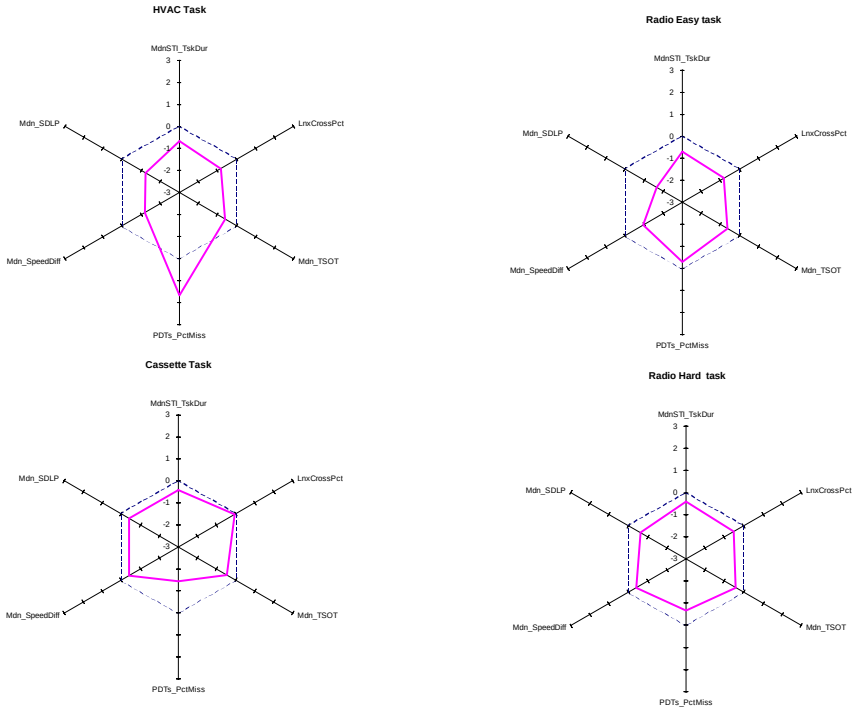
Selected Star Plots: “Higher Workload” Visual-Manual Tasks



Legend			
Label	**Description**	**Label**	**Description**
Mdn_SDLP	Median Standard deviation of lane position	MdnSTI_TaskDur	Median Task Duration in STISIM
LnxCrossPct	Percent of trials with a cross of lane line	PDTs_PctMiss	Peripheral Detection Task in STISIM Percent Missed Detections
Mdn_SpeedDiff	Median Speed difference	Mdn_TSOT	Median Total Shutter Open Time

Figure 8-4. Star Plots of Selected Laboratory Surrogate Measures: Higher-Workload Visual-Manual tasks

Star Plots for Laboratory Surrogates:
"Lower Workload" Visual-Manual Tasks



Legend			
Label	**Description**	**Label**	**Description**
Mdn_SDLP	Median Standard deviation of lane position	MdnSTI_TaskDur	Median Task Duration in STISIM
LnxCrossPct	Percent of trials with a cross of lane line	PDTs_PctMiss	Peripheral Detection Task in STISIM Percent Missed Detections
Mdn_SpeedDiff	Median Speed difference	Mdn_TSOT	Median Total Shutter Open Time

Figure 8-5. Star Plots for Selected Laboratory Surrogate Measures: Lower Workload Visual-manual Tasks

Now consider star plots for the lower-workload tasks. HVAC is unusual in that it has surrogate values well below the mean value of zero on each radial for the task set except for the PDTS Percent Miss Rates. This reflects the paradox of short tasks associated with poorer OED performance. All of the other tasks are below the mean value of zero represented by the annulus ring. But each task does have unique profiles. For example, the Radio (Easy) task had especially low SDLP and Speed Difference scores. On the other hand, Insert Cassette had Percent Lane Exceedance (Cross) trials scores approaching those of Manual Dial.

This type of multivariate view provides insights into the nature of the task impacts on various surrogate measures of workload. It may also be a means by which to assess tasks in the future. For example, if key dimensions are selected for the star plot, these may support a non-compensatory decision rule (Gigenrenzer, Todd, and the ABC Research Group, 2000). A non-compensatory decision rule is one in which a good score on one variable or dimension (e.g., SDLP) does not offset a poor score on another variable or dimension (e.g., OED Miss Rate). The star plot might provide a graphical tool to apply this rule. If a task has any arm beyond the normal or other acceptable annulus or pattern, then it would be a candidate for further revision. In the star plots above, the annulus used for reference is simply the mean of all tasks in the set, but it could be used to graphically illustrate a criterion level on each radial to which tasks would be compared in order to be judged acceptable.

8.3 Other Issues in Driver Workload Assessment

8.3.1 Selected Research Hypotheses and Their Validity

In Chapter 2, Table 2-3 presented research hypotheses for selected driving performance measures and selected task-related driver eyeglance measures. Chapter 2, Table 2-4 summarized research hypotheses for selected laboratory surrogate performance measures and also subjective assessments. Generally speaking, directional research hypotheses were tested by the rationale presented in Chapter 1. There are many criticisms that can be leveled at the hypotheses in these tables, particularly Table 2-3. Two classes of these criticisms will be addressed below.

8.3.1.1 Dissociation and Driver Workload

One class of criticism for the directional hypotheses in Table 2-3 is based on the concept of dissociation from the driving task. Because of dissociation, driver steering and accelerator input measures may be more sensitive than those measures listed in Table 2-3. For example, drivers under subsidiary task load may stop modulating steering inputs. The idea is that normal driving is characterized by small, relatively uniform steering corrections to maintain lane position. As task demand increases, the small corrections cease and the steering wheel is held constant for a period of time. The steering hold is usually accompanied by increased variability in the time the steering wheel velocity is zero or by large or high-velocity steering corrections. Such measures have been applied to in-vehicle tasks by Dingus, Antin, Hulse, and Wierwille (1989) and Nakayama, Futami, Nakamura, and Boer (1999), and others. Greater sensitivity to task differences might come from measures of driver input. However, the relevance of such measures to safety is less clear than measures of vehicle response, i.e., position and velocity. For example, steering input is two time derivatives removed from vehicle lateral position.

The dissociation concept can be moved up to the level of vehicle position and velocity and apparently turn the directional hypotheses in Table 2-3 on their heads. The idea is still that a driver under high task demand will dissociate from the driving task. This effect, if present, would result in less rather than more lane position variability, smaller rather than larger speed differences, fewer rather than more lane exceeds, etc. This “frozen in distraction” hypothesis is worthy of further research to determine if, when, and how it applies. Until that time, pragmatic considerations can turn the Table 2-3 research hypotheses upright again. From the standpoint of other road users, it does not matter what the level of distraction supposedly is, as long as the distracted driver remains well within the lane, keeps adequate separation from the vehicle ahead of it, does not vary much in speed to vex a following driver, and reacts adequately to objects and events. If any of these conditions are not met, the Table 2-3 directional research hypotheses re-enter the discussion.

8.3.1.2 Driver Workload and Compensatory Driving Strategies

Another alternative view is that some workload measures reflect compensatory driving strategies in the face of distraction rather than indications of disrupted driving. Lane exceeds might be interpreted, in this alternative view, as driver choices to provide a wide berth to the left, or perhaps to the right. Similarly, speed variability associated with falling back or slowing might be interpreted as the means by which the distracted driver provides himself or herself with an added safety margin from the vehicle(s) ahead. As a last example, increased standard deviation of lane position might be interpreted as relaxation of the driver's criteria for lanekeeping performance that he or she considered appropriate during that task.

Consider the effect of speed variability or falling back as a compensatory strategy. It seems on the face of it that this would be a good way to increase the safety margin between the distracted driver and the vehicle(s) ahead. But this behavior has its potential costs as well. Shinar (1998) recently completed a critical review of the effects of speed and speed variability on traffic safety and crash incidence. While the data are incomplete, the main trend in the research is that speed variability is more hazardous than speeding per se. That is, speed variability, defined as the difference between subject vehicle speed and prevailing traffic speed, is more likely to contribute to a crash than is, say, all vehicles traveling in excess of the posted speed limit but within a narrower range. Beyond this, increasing time headways between vehicles increases the probability of a cut-in as another vehicle changes lanes to take up the space (Fancher et al., 1998). Effectively, cut-ins reduce the safety margin supposedly gained by falling back.

8.3.2 The Role of Task Duration in Workload Measurement

Task duration can limit the maximum values various driving performance measures can attain. This is largely due to the physical nature of the vehicle. For example, coast-down for a passenger car takes a certain amount of time to go from 55 mph to 45 mph. Tasks shorter than that period of time will show less speed difference than longer tasks. Similarly, given a constant steering wheel angle, lane position will vary less over shorter periods than over longer periods of time.

Task duration is sometimes inherent in the task itself. The 13 DWM visual-manual tasks were of this nature. Such tasks as radio tuning, map search, HVAC adjustments, and destination entry, each are defined in terms of goals and criteria that operationally define the end of the task. The task takes as long as the driver needs to reach those goals or criteria. These criteria provide natural constraints on how long the driver will be engaged in the task while concurrently driving. In this sense, task duration is an essential attribute of the task. Time-constrained driving performance and eyeglance measures should be interpretable as aspects of workload in such cases.

In other DWM tasks, the duration of the task period is arbitrary. The duration needed to insert a CD from a visor wallet and select Track 7 is constrained by the equipment design and the driver's efforts. But what is the natural duration of listening to a book on tape, engaging in dialogue, or just driving? There is no answer to this. Procedurally, auditory-vocal tasks were arbitrarily set to two minutes in length. Visual-manual tasks took as long as needed to reach the goal state for each. There was also at least one “crossover” task for each category. Book-on-Tape Summarize was a short auditory-vocal task and Destination Entry was a long visual-manual task. In general, the effects of time on driving performance measures were consistent. Longer tasks showed more variability in lateral and longitudinal control than shorter tasks. Shorter tasks paradoxically showed poorer OED performance than longer tasks. Overall, the effects of non-arbitrary task durations should be considered findings, not artifacts. On the other hand, arbitrarily set task periods may be artifacts rather than findings. One way to remove the ambiguity may be to set tasks to be compared to the same duration, if possible. This may not always be feasible, however, as was discussed previously.

8.3.3 Why Not Include Failure Trials in a Workload Assessment?

Obviously, task failures should not be ignored. Task failures can indicate high workload, especially if the failure occurred despite attempts at error recovery. But interpretation is complicated when the performer does not successfully complete a task. If the goal is not reached, the person may persist, creating an outlier on the high end of task duration or another workload measure. Alternatively, if the goal is not reached quickly, the person may decide to quickly abandon the task and create an outlier on the low end of the same measures. It is also possible that the failure was caused by a speed-accuracy tradeoff. Either way, a task failure can increase a measure's variability in opposite directions.

In fairness, successfully completed tasks are also uncertain in workload assessment. A person may have guessed correctly; been very familiar with task content (e.g., the subject of a text message); had unusual cognitive abilities (e.g., a bookkeeper's mental arithmetic skills); and so on. Experimental procedures like counterbalancing and random selection are intended to address such problems. This is not the case for issues raised in the previous paragraph.

Task failures can reflect usability problems. Usability problems could be addressed independently of workload assessments and preferably before. Usability problems can increase driver workload. But usability does not guarantee low workload. A task may be designed for ease of use and yet be composed of many task steps that are still too much to pursue while driving.

8.3.4 Detection Versus Response in Object-and-Event Detection Trials

The term detection is an interpretation applied to response data. This is another simplification that may be wrong. A test participant might have detected something but did not respond in time. The method used in the DWM research required all responses to be made within the task period. An analysis was carried out to look for responses up to five seconds beyond the end of the task. This analysis showed that the shortest tasks (largely visual-manual tasks) did have more responses given after the task was finished. The effect diminished as task length grew. These additional detections were nevertheless not included in the detect data. This decision was made because of uncertainty with regard to their meaningfulness and the fact that they did not fundamentally change the duration-related paradox. Multiple Resource Theory (MRT) predicts that manual task load (device holding, button pressing, knob turning, map tracing) can interfere with another manual response. The “too late, not counted” policy is, therefore, compatible with a workload interpretation.

A test participant might also have responded without detection before he or she completed a task. This can be considered a “when in doubt, shout” policy. This type of problem is not easily assessed in the DWM data. Research experience suggests that test participants usually try to comply with experimenter instructions. Appropriate caution in interpretation is assumed.

8.3.5 Why Some Driving Performance Metrics Seem More Interpretable than Others

Range is defined as the distance between subject and lead vehicles. Generally, it would seem that smaller range would be indicative of less safety margin to recover from unexpected events. However, there are several methodological reasons why this measure cannot be readily interpreted in this study. The test participants were prompted between trials to adjust the range or separation to the lead vehicles to achieve approximately equal initial conditions for the next trial. This also resulted in a subject vehicle travel speed adjusted to a lead vehicle speed of approximately 55 mph. These “mandated” initial ranges may not resemble self-selected vehicle separations that a driver might select. Beyond this, measures like mean range or average speed are hard to interpret. Assume that the speed or separation measures reflect task duration. Shorter tasks will appear to be associated with closer car following and higher average speed than longer

tasks that involved more falling back behavior due to the higher workload. Because the experimenter effectively placed the driver in the initial conditions of range and speed, these results are artifacts of the experimental method.

Measures of lanekeeping variability and speed difference are not subject to such artifacts. There are several reasons for this. First, variation or difference is by definition zero at time zero (task start). Second, there may be higher-order variables like lateral velocity and acceleration that are generally non-zero at task start. However, these were not systematically manipulated like initial vehicle time headway (defined by both range and speed). Third, task duration can constrain the maximum achievable variation or difference values observed. But it is always possible to minimize variation regardless of task duration (though longer tasks require more effort).

8.3.6 Correlations and R-Squared

Correlation and linear regression have been presented to address the issue of both repeatability and predictive validity. This is a useful simplification. Like all simplifications, several limitations should be kept in mind (Netter, Kutner, Nachtsheim, and Wasserman, 1996). A high correlation coefficient does not necessarily indicate that useful prediction can be made. It depends on the precision required. It is easier to make ordinal predictions into higher-workload and lower-workload categories than to make fine gradations between tasks. A high correlation does not necessarily mean the regression line is a good fit. Extreme points can torque a regression line in misleading ways. Conversely, a low correlation may simply reflect the presence of a curvilinear relation rather than a linear one. This is why bivariate plots are useful. A high correlation does not mean that it will be as high in a new sample. Regression analysis cannot, in itself, define causal relationships. Prediction depends on the similarity between the range of observations initially modeled and new applications.

8.3.7 Why Performance Might Improve With Task Load

The Yerkes-Dodson law states that performance is best at intermediate levels of arousal or stress (Fitts and Posner, 1960; Wickens and Hollands, 2000). Low arousal or high arousal can both reduce performance relative to the middle range. According to this law, a plot of performance (the y-axis) as a function of arousal level (the x-axis) resembles an inverted “U”. A plausible hypothesis is that some level of task load improves (otherwise boring or monotonous) driving performance in accord with the Yerkes-Dodson Law. Such an effect merits further research into optimal levels of task loading.

8.3.8 Cautions Regarding Venue Differences

Although the DWM study focused on development of performance metrics and not tasks per se, one important finding has relevance to driver workload research conducted by others. In this study, there were effects observed in the laboratory that were not observed on the road. For example, no object and event detection metrics discriminated (hypothesized) high- and low-workload for auditory-vocal tasks on the road, while some laboratory surrogates did. Possible explanations for this are that drivers may perceive risks differently in the laboratory than in real vehicles. Or perhaps the on-road experiment was somehow insensitive to these effects in an uncharacteristic way. Until this discrepancy is better understood, judgments on task effects should not be based solely on laboratory results. Nonetheless, surrogates can be used iteratively through product development to manage workload implications of new system designs subsequently verified by on-road evaluations. A recommended “toolkit” will be defined in Section 8.5 to support this process. The toolkit includes both non-driving and driving-based tools.

8.4 Discussion of Findings Across Multiple Measures of Driving Performance

The purpose of examining correlations between the driving performance metrics, and of constructing star charts from standardized data on those metrics, was to gain insight into any patterns of interference with driving performance during multitasking that may emerge across measures. It was recognized, in studying these interrelationships, that the construct being studied (driver workload-caused distraction or interference with driving) was likely to be multi-dimensional in nature, meaning that it was thought to be:

- Represented in the data by simultaneous effects on multiple variables,
- Reflective of allocations of driver resources across input modalities, output modalities, working memory, and central attention,
- Reflective of allocations of driver resources across multiple activities and allocations of attention across spatial areas of interest (central forward road, peripheral areas including mirrors, and inside-vehicle), and
- Reflective of adaptive driver strategies that were responsive to demands and dynamically varying across time.

Vision is the primary driver input resource. The primacy of vision in driving necessitates an emphasis on eyeglance measures. It is of special interest to see how eyeglance measures relate to other aspects of driving and object and event detection. This emphasis is reflected in the discussion provided below.

8.4.1 What the Relationships Within the Data Say About Workload/Distraction

The relationships between driving performance variables yielded the following insights about driver workload and distraction:

- States of driver workload that produced overload or interference with driving performance were manifest not on just one underlying dimension of performance, but on several simultaneously affected ones, confirming that workload-induced distraction is multidimensional in nature. Driver underload, though not a focus of this research, might be countered by some subsidiary task activity.
- There were specific patterns of effects that appeared to be interpretable as characteristic of distraction, but which will require further research to confirm.
- Different patterns of interference/degradation across the categories of performance were associated with, and diagnostic of, different types of distraction.
    - Different patterns emerged for different types of tasks (auditory-vocal, Just Drive, and visual-manual tasks)
    - Different patterns emerged for individual tasks, that were unique to the demands that each imposes on drivers

An interesting observation about this study relates to the use of the three-car platoon approach (in which the driver had both a lead car in front and a follow car behind). This approach was selected in order to represent scenarios that were identified during Task 1 as important among those giving rise to distraction-related crashes. And indeed, a more recent analysis confirmed this. Based on

estimates from the 2000 GES database, Foley, Glassco, Cohen and Chang (2004), provided a table showing that the number of crashes annually attributed to distraction (of all forms) totals approximately 748,000, and that 67 percent of these are typically longitudinal crashes which involve a lead vehicle that is decelerating, stopped, or moving. Given that the DWM experiment conducted on the road and on the test track included a car-following scenario with event detection stimuli (one that was a lead-vehicle "coast-down" deceleration without brake-light illumination, and one that consisted of CHMSL illumination, but without the usually-associated deceleration so as to approximate "riding the brakes"), it allowed exploration of effects not previously available for study during task-engagement and driving. A number of driving performance measures, including glances to the road and mirrors, measures of event detection, and to some extent speed-keeping, carried information of interest for understanding the effects of different task types on driver attentiveness to driving.

Nonetheless, although the scenario used in this experimentally-controlled work was relevant for conditions under which distraction sometimes arises, it is important to recognize that the data from this study are not a sufficient basis from which to begin estimating crash risk in more naturalistic driving environments. First, data on many more contributing factors (e.g., the occurrence of events in the real world, such as sudden vehicle stops and especially the probability with which such events co-occur in the real world with task use by another driver, would need to be known to even begin relating task effects in the experimentally-controlled conditions of this study to crash risk in a naturalistic driving environment. This study, as previously discussed, focused not on the relationships between in-vehicle device use and crash risk, which would require information beyond that acquired in the scope of this study. Rather, this study focused on the relationship between in-vehicle task performance, the demands imposed on the driver, and any degradation or interference with driving that resulted. Furthermore, it is important to keep in mind that the data have some limitations. Drivers in the study knew that special precautions had been taken to minimize the likelihood of a crash during the experiment, and this may have in some way altered their behavior (e.g., their willingness to look away from the road, their choices of how long to look away and when to look away). Until replications are done under fully-naturalistic driving conditions, it will not be known whether the glance patterns observed in this study generalize beyond the experimental conditions under which they were observed.

Another noteworthy thing about the findings was the prominence of event detection findings in it. The process of detecting and responding to events appears to be an integral part of driving that fundamentally involves strategies for scanning the forward road, relevant periphery (including mirrors), and monitoring speed, headway, and lane position. And the event detection interactions with eyeglance behavior suggest that there is not a single strategy reflected in the data, but probably a variety of dynamically changing behaviors that differ with the type of event that is occurring on the road, as well as with conditions of workload being carried by the driver, and other variables. This is evident in the fact that glance durations change differently depending on event type, once an event has been detected, and also apparently with amount of traffic, since there were some differences between road and track findings. Similarly, some event types led to concomitant changes in other behaviors, such as increased scanning of mirrors, which may be consistent with the finding reported in the literature that situation awareness heightens in salient areas when a threat has been detected (cf. Gawron, 2000). Also very interesting, the information about these changes was not all carried in a single predictor variable.

Event detection is difficult to study because often the methods used to study it alter the behavior of interest, and it has, as a consequence, been studied less than many other aspects of driving performance. However, its integral role in driving (underscored in the multiple effects observed here) suggest that it holds many insights into the phenomena associated with driver workload and distraction, and really needs to be tackled more vigorously in future research efforts.

Finally, it is important that key information about glance behavior is not all carried in traditionally-used variables like the eyes-off-road-time that is associated with task performance or number of glances to the task. Nor is it sufficient to measure simply eyes-on-road and eyes-off-road. Significant information is also being carried in glances to other locations. These analyses have demonstrated that very important information is contained in glance metrics associated with other location types (e.g., glance durations to the road, number and durations of glances to the mirrors). This information is especially relevant for auditory-vocal tasks, which showed different glance patterns than visual-manual tasks, and for event-detection. This finding, too, should be of benefit to future research programs.

One of the most striking findings from the study emerged from the analysis of eyeglance data together with event detection data. In this analysis, duration of glances emerged as much more important in a diagnostic sense than had been expected. The detection of events by drivers changed their glance patterns, and duration changes appear to have played a central role in the altered scanning patterns. Detection of events had a significant effect on glance durations, and also, therefore, on number of glances and rate of glancing. For the CHMSL and FVTS events, durations of glances to the road decreased slightly, but for LVDs, durations to road increased in the test track data, and different patterns for the road. There were changes to other glance durations as well. For example, maximum glance duration emerged with significant interactions in the Linear Mixed-Models analysis of the eye data when event detection was included as a variable. When both the eye data and event-detection data were considered together, the duration changes seem to underlie many of the observed patterns on multiple metrics. Furthermore, changes to glance durations interacted with task type and were more pronounced for Just Drive and auditory-vocal tasks than for visual-manual tasks, which usually showed a different pattern. These findings suggest that there is much more to understand about the role of glance duration in glance behavior during driving—and its role has perhaps been underestimated in the past. Glance duration may be one key to understanding the underlying processes governing scanning—through which the driver determines where to glance next and for how long. Some of these processes may occur outside of conscious awareness and some may occur within conscious awareness. As such, both brain imaging and behavioral science approaches may be needed in order to push the state of understanding forward in this arena.

8.4.1.1 Effects of Multitasking Are Multidimensional and Specific to Its Structural Interference With Driving

The findings from the study are consistent with the notion that interference between a task and driving is specific to the demands it places on driver resources, and the ways in which those demands conflict with the competing demands placed by driving and event detection (of various types).

This can be seen, for example, from the star charts that depict an average visual-manual task and an average auditory-vocal task plotted in Chapter 3, Figure 3-58. These suggested that when an in-vehicle task was concurrently performed while driving, not just one dimension of performance was affected, but multiple dimensions of performance were affected. Furthermore, individual task plots (shown in Appendix S) indicated that patterns of interference with driving performance were also not unidimensional, but multidimensional and highly specific to the task, with some similarities between task types. The fact that patterns of interference are task-specific carries important information about the constructs of workload and distraction.

The star charts are consistent with findings in the published literature that support this notion. For example, research has shown that the extent to which an in-vehicle task interferes with the ability to perform the driving task depends crucially on the maneuver underway (indicating specificity in the interference between in-vehicle task and driving demands). Verwey (1991) reported a study

showing that driving situations differ in the extent to which they are demanding of attention. In particular, different driving maneuvers require different information processing and attentional resources. Performance on a in-vehicle task, such as auditory or visual serial addition, suffered more when drivers carried it out while doing a turning maneuver than while driving straight, for example. Furthermore, some maneuvers required more visual processing than others, and thus were interfered with more by concurrent performance of a visual secondary task. Duncan, Williams, Nimro-Smith, and Brown (1992) examined driving performance measures while drivers concurrently said aloud a single digit per second that they were instructed should be unrelated to the previous digit. They found that some elements of driving performance were influenced by concurrent performance of the in-vehicle task but not others. For example, during random digit production, drivers applied their brakes later when approaching intersections and tended to check mirrors more but at inappropriate times. However, other measures of driving performance were unaffected by the concurrent task. These findings of differential task interference, and others, are not consistent with the notion that a driver's central attentional capacity is exceeded whenever two tasks are performed simultaneously. Rather, the results are indicative of specific interference between tasks when the tasks simultaneously demand use of the same or similar perceptual, information-processing, or response resources.

Further, driving has sometimes been characterized as largely automatic, implying that there is little or no cost to supervisory attentional processes (Groeger, 2000). This leads to the expectation that multitasking during driving can be done with no reduction in performance. However, several studies have demonstrated that this is not the case. Groeger and Clegg (1998) and Shinar, Meir, and Ben-Shoam (1998) have shown that highly practiced processes, such as gear changing, do require attention (though not necessarily conscious attention). In the Shinar et al. (1998) study, drivers drove either an automatic or manual shift car over a fixed route. They were asked to signal when they detected either of two types of road signs. Drivers using automatic transmission vehicles rather than manual shift vehicles correctly detected more target signs, suggesting that manual shifting diverted more attention than did automatic shifting. If gear shifting were so highly practiced that it required virtually no attentional resources, then performance with manual shift cars should have been indistinguishable from performance with automatic transmission vehicles. This is not what occurred. Furthermore, while differences between novice and experienced drivers were observed, both groups were affected by the attentional demand of shifting gears to the point that sign detection was affected.

Findings such as these from the literature together with the findings from the DWM study, particularly as depicted in the individual task plots, lead to the conclusion that interference between tasks is predicted not just by their type (auditory-vocal versus visual-manual) or by their difficulty per se, but by the structural overlaps between the resource demands of the two tasks (Groeger, 2000). This is a central tenet of Multiple Resource Theory (MRT) (Wickens, 1980) and supported the use of the Wickens' framework in the DWM project (see Appendix A, *Rationale for Selecting Tasks*). MRT provides a way to identify the structural interference between driving and other concurrently performed tasks, and suggests that time-sharing difficulties between tasks arise under conditions in which structural interference occurs. It is hypothesized that driver inattention and distraction are examples of such time-sharing difficulties, which arise when structural interference occurs between concurrently performed tasks. While the specific formulation of the Modified MRT model implemented here did not prove to have as much predictive power as hoped, continuing development of the underlying theory is warranted on the basis of the empirical findings related to the observed specificity of interference between driving and tasks.

8.4.1.2 Task Types Do Differ in Their Effects on Driving Performance and Event Detection

The differing effects of task types on driving performance are apparent in several ways from the results of the analyses of driving performance.

Eyeglance metrics showed distinct patterns for different types of task engagement (just driving versus concurrently performing an auditory-vocal task or concurrently performing a visual-manual task). Effects observed on the road for in-vehicle tasks on eyeglance behavior replicated those observed on the test track. Vehicle control metrics also showed different patterns between visual-manual and auditory-vocal task types, although less distinct. Based on the road data, a summary of the profiles follows.

The Just Drive task was distinguished by patterns in which drivers looked at the road about 81 percent of the time and scanned their mirrors about 15.4 percent of the time. Glances on the road were about 3.8 seconds duration, on average.

Auditory-vocal tasks showed a similar pattern, though drivers gazed at the forward roadway somewhat more (87%), using longer gazes (4 to 8 seconds, on average), and scanned their mirror somewhat less (11.4%). This miss rate for event detection was slightly elevated over just driving for auditory-vocal tasks for CHMSL and LVD events, showing an increase of ~6 percent for CHMSL and ~4 percent for LVD events; and somewhat more for peripheral FVTS, an increase of ~14 percent, though event detection was less affected by auditory-vocal tasks than by visual-manual tasks. When events were detected and responded to by drivers, scan patterns subsequently changed in a way that appeared adaptive to the specific type of event to which the driver had responded, and the nature of the response was such that further detection of subsequent events (had there been any) would have been higher. This led to the conclusion that event-detection may serve as an "attentional interrupt" for auditory-vocal tasks and the task of just driving, resulting in more active scanning of the road and mirrors for situational awareness. In terms of lanekeeping, median SDLP was similar to and slightly smaller, on average, than Just Drive (and not discriminable from it). In terms of speedkeeping, auditory-vocal tasks led to slightly larger speed differences than Just Drive (but these were not discriminable from it in formal analyses of discriminability).

Visual-manual tasks showed a different pattern, in which drivers looked at the forward roadway much less (viewing the road only 42% to 68% of the time during a task), and using glance durations on the road that were much shorter (less than two seconds long, on average). This reduction in glances to the road was made in order to view task-related areas required for performing the in-vehicle activity (viewing the task 24% to 52% of the time during its length). For visual-manual tasks, glances tended to cycle frequently back-and-forth between the task and the roadway locations, and glance-rate measures proved to carry interesting information. Visual-manual tasks led to a more pronounced reduction in mirror-scanning (to 6.4%) and were associated with higher rates of missed events (though this was sometimes due to a methodological constraint for LVDs). Increases in miss rates over Just Drive were approximately 23 percent for CHMSLs, 28 percent for LVDs, and 65 percent for FVTS events on average.

When events were detected and responded to by drivers, scan patterns subsequently changed in a way that appeared specific to the type of event to which the driver had responded. Furthermore, the changes had different implications for visual-manual tasks than for auditory-vocal tasks. However, for visual-manual tasks, more active scanning of the road and mirrors following detection of an event appeared to occur only for LVD events. For other event types, when an event was detected during a visual-manual task, visual scanning between task, road, and mirror locations seemed to increase (apparently without task shedding). High glance rates appeared to be related to higher miss rates for these events, so this elevated scanning rate had an adverse

consequence for visual-manual tasks (due perhaps to the fact that the eyes are in transition more of the time since there are more locations being transitioned among, and hence more time during which vision is suppressed, and perhaps events missed). In terms of lanekeeping, effects were observed for the two most difficult visual-manual tasks on median SDLP and Percent Trials with a Cross of the Lane Line. However, median SDLP averaged (across all visual-manual tasks) less than Just Drive, and was not discriminable from Just Drive. In terms of speedkeeping, small differences in speed were observed during visual-manual tasks, and Speed Difference tended to be less than for Just Drive. These effects are correlated with task duration in ways that suggest how long it takes to complete a task is an important part of workload assessment.

8.4.1.3 Visual-Manual Distraction Effects are Larger in Magnitude

Visual-manual tasks had a more pronounced effect on driving performance than the auditory-vocal tasks. This is illustrated in Figure 8-6 below for the metrics (based on the test track data):

- percent of task spent looking at the road;
- percent of task spent looking at the mirrors;
- percent of CHMSL events detected; and
- percent of FVTS events detected.

For each metric in the figure, the Just Drive task is plotted along with the average auditory-vocal task and the average visual-manual task. For the percent of task spent looking at the road (upper-left graph), the auditory-vocal tasks involved approximately 5 percent more time on-road compared with Just Drive, while the visual-manual tasks were associated with about 40 percent less time on the road. In this case, the magnitude of the visual-manual task effect is eight times larger than the auditory-vocal task effect. More subtle differences are depicted for the percent of task looking at mirrors metric (upper right graph). Here, the difference between Just Drive and auditory-vocal tasks is 3 percent versus 6 percent for the Just Drive and visual-manual tasks comparison. The percent detect CHMSL and FVTS events show similar results in that the magnitude of the auditory-vocal effects are smaller than the visual-manual task effects. Another possibility for visual manual tasks to show larger magnitude for these metrics is because visual manual tasks by nature were head-down and had shorter durations as compared with auditory vocal tasks.

Visual-Manual Tasks had More Pronounced Effect on Driving Performance Trials

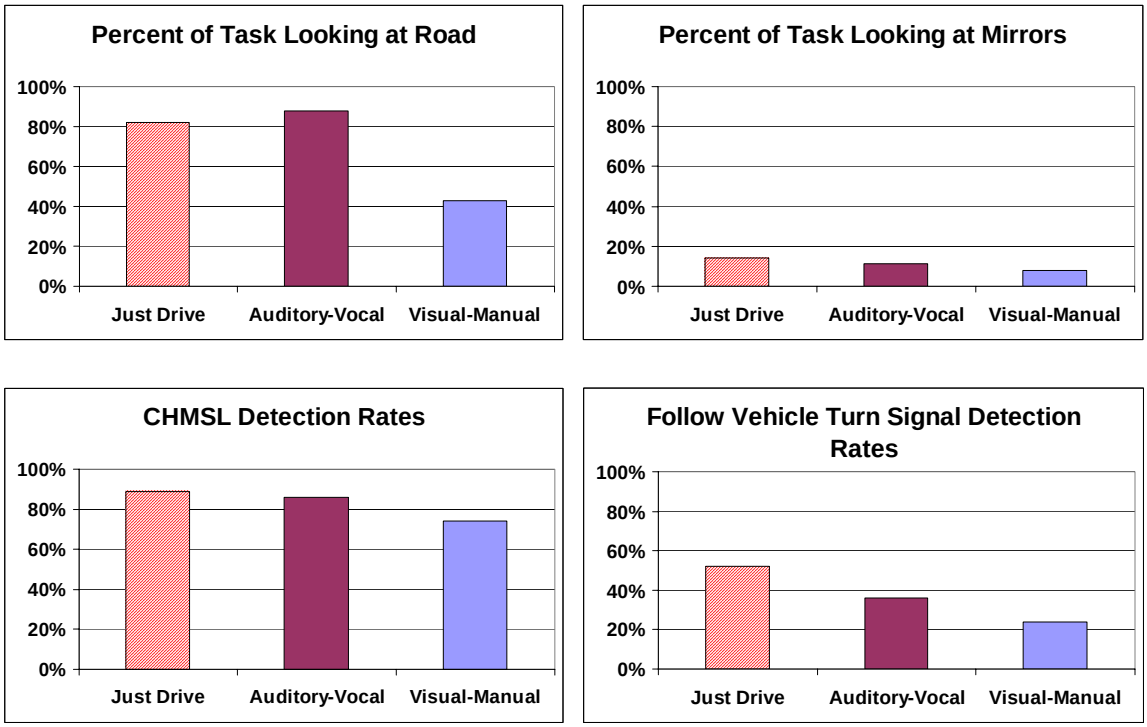


Figure 8-6. Comparison of the Magnitudes of Visual-Manual Task Effects With Auditory-Vocal Tasks

8.4.1.4 Cognitive Distraction Effects Appear Very Subtle

Identifying any measurable interfering effects of cognitive load on driving performance has proven very difficult, in part due to the fact that it is difficult to devise tasks that impose cognitive-only (or even primarily cognitive) loads on the driver. This is necessary to clearly discern the effects of cognitive load on performance. Usually cognitive load co-occurs with other types of task loading (visual, manual, auditory, vocal) so it is not always clear whether any observed effects on driving performance can be attributed to the cognitive portion of the loading. However, a more important issue is that when the visual and manual loads are eliminated to help isolate the cognitive operations and a task allows the eyes to be forward, on the road and/or scanning the traffic environment, the magnitude of intrusions on event detection, as compared with visual-manual tasks, is smaller and more similar to just driving, as are the magnitude of effects on lanekeeping and speed variability, though some of these effects are discriminable from just driving. A clear interpretation of whether there is an intrusion on driving from cognitive load is thus very difficult.

Nonetheless, several effects in the auditory-vocal correlations appear to be related to cognitive load, though again these effects were in most cases small relative to the full range of differences among all tasks and in some instances accounted for little variance. The following summarizes the effects, where effects were found. Cognitive load appeared to lead to a sort of cognitive tunneling effect in which gaze was directed at the forward road, scanning of the mirrors was reduced somewhat (though to a lesser extent than for visual-manual tasks) and some peripheral events were missed (though to a lesser extent than for visual-manual tasks). Cognitive load

appeared to lead to a few (usually only one or two) upward glances during the auditory-vocal tasks that involved working memory operations and/or language production (sometimes associated with habits of looking at a listener). These few task-related glances were many fewer than those related to visual-manual tasks.

The relationships between variables that may possibly be related to cognitive load offered some support that was consistent with preliminary findings reported from HASTE, findings reported to be in use at Volvo as a means of diagnosing cognitive distraction, and findings reported to be from Saab for diagnosing attentional focus. However, these factors also contained information which conflicts with the European interpretations of data and hypothesized relationships, as explained more fully in what follows.

In a 2004 paper entitled, *Driving Support From Visual Behavior Recognition (VISREC) – Evaluations of Real-time Attention Support Functionality*, given at the International Workshop on Progress and Future Directions of Adaptive Driver Assistance Research (Victor, 2004), described a system developed to adapt to driver workload that was in some way calibrated to eye movements, using a metric of Percent Road Center, based on the finding that drivers look 95 percent of the time to their future path. He referred to a Saab finding that 80 percent of the driver's looks are to this region. He furthermore reported work underway to develop a Cognitive Distraction Alert wherein the system determines cognitive distraction using an algorithm based on Percent Road Center gazes. It was hypothesized (based in part on preliminary HASTE findings) that there were:

- Increased spatial gaze concentration on road center during cognitive distraction
- Improved lateral control (lanekeeping)
- Hypothesized reduced event detection (but neither Volvo nor HASTE had data on event detection)
- Regan and Gray (2000) were cited as describing judgment errors that stemmed from looking too much at road center, causing extra RT
- Recartes and Nunes (2003) were also cited as well as Victor and Karlsson (in preparation) regarding foveal versus peripheral effects

The findings in the DWM study are consistent with some but not all of these European findings or hypothesized effects. The DWM data do demonstrate long look times to the forward road for some tasks in some conditions, primarily for auditory-vocal tasks. The univariate analysis of Proportion of Task Time Spent Looking at the Road is an excellent example of this. The metric of Mean PctDurRD, or the Percent of Task Time Spent Looking at the Road during an auditory-vocal task, did show high negative correlations with SDLP and Speed Difference (indicating that standard deviation of lane position and speed differences decreased as more time during task was spent looking at the road, consistent with the European hypotheses). However, correlations between other measures in the eye data and other driving performance data reveal that concentration of gaze on the forward roadway is NOT necessarily associated with improved lateral control or with reduced event detection. For example, total glance time to the road during the task, and glance durations on the road showed high positive correlations with SDLP and Speed Difference metrics (an outcome that appears inconsistent with the European hypotheses). It should be noted that it was not possible, however, to compute the exact measures used by Saab and by Volvo, which apparently require knowing locations of gaze with more precision than was known in this study, namely, within a few degrees or perhaps even arc minutes of road center, so precise comparisons with the European measures could not be made, only approximate comparisons. SDLP covaried with glance durations to the road in much the same way as Speed Difference did. These variables appeared to be positively related to hypothesized cognitive load

in some instances and negatively in others, perhaps a more complex relationship than the one observed/hypothesized in the early HASTE findings as reported by Victor (2004). Also, the DWM project data revealed that gaze concentration on the road (if measured in terms of glance duration to the road, length of glance durations on the road, and total glance time to the road) is sometimes (but not always) associated with reduced event detection, and depends upon the type of event that occurs.

The DWM project findings would thus seem to suggest that real-time adaptive attentional support in future vehicles may be far more complex to achieve than the operation of initial systems on the market currently reflect.

Furthermore, there is evidence in the DWM project data on auditory-vocal tasks that working memory plays a role in multitasking while driving, and interestingly, may help explain what it is that allows drivers who are performing auditory-vocal tasks to respond so fully to events when they detect them.

Recent research has suggested that working memory and attention play a key part in distraction. In fact, some of the early research on this subject (see Figure 8-7) suggested that task demands competing for the structural resource of working memory may interfere with the ability of supervisory attention to maintain focus and priority, as well as to inhibit responses to stimuli that are not relevant to a task being done. In a subsequent article, Lavie (2005) took these findings a step further, and proposed a notion of task-loading that brings together research on attention and executive brain function. It is based on research indicating that a high perceptual load (early in processing) that engages full attention prevents interference from distractor stimuli, but high task loads on cognitive and working memory functions interfere with the ability to maintain priority and focus, and thus increase interference from distractor stimuli. The terminology of distractor stimuli stems from the experimental paradigm used by Lavie and colleagues.

However, if the terminology is changed slightly, it may be that their work can help explain some of the findings observed in the DWM study on detection and response to events. If the events in our study can be thought of as distractors in Lavie's terms (that is, as stimuli that are outside the focus of the activities being actively multi-tasked on a continuous basis, which might be hypothesized to include lanekeeping, speedkeeping, and an in-vehicle activity), then when the in-vehicle activity (coupled with the visual monitoring of the roadway ahead and mirror scanning) imposes a high perceptual load, all perceptual capacity would be used up, and an event which occurs may not be perceived. Of course, it may also not be viewed, if the eyes are off the road and on the device, or if the eyes are in transition between locations. But according to the work described by Lavie, even if fixated by the eyes, it will not be perceived if the perceptual load has already fully used visual attention to capacity).

On the other hand, if perceptual load is low to moderate (say with an auditory-vocal task), events occurring on the road have a higher likelihood of being seen and perceptually processed. However, if the task is imposing a cognitive load on working memory, it may interfere with the ability of executive attention to maintain focus on the auditory-vocal task (to the exclusion of unexpected stimuli), and may also interfere with the ability of executive attention to inhibit automatic response processing that may be triggered by an event which occurs. This may benefit the driver in some ways. For example, the driver may respond to a detected event fully and more often during an auditory-vocal task that imposes a cognitive load, than during a visual-manual task that imposes a high perceptual load (perhaps because executive attention is unable to inhibit responses to stimuli outside of the in-vehicle task).

Example Study Described in Lavie (2005)

One such study appeared in Science (de Fockert, Rees, Frith and Lavie, 2001) entitled, "The Role Of Working Memory In Visual Selective Attention." The authors reported research in which subjects performed a task requiring that five digits be held in working memory. Subjects had to remember either a fixed order of five digits (e.g., 01234) or a different order of digits on each trial (e.g., 03124) (with memory probes to ensure that information was being held in working memory). While remembering the digits, subjects were also asked to perform the following task. They were to view a photograph of a celebrity, followed by the name of a celebrity (shown in text). Sometimes the named celebrity was the same as the one in the photograph and sometimes it was different. Subjects were to respond by categorizing the celebrity named in the text presentation as either a politician or pop star. (This essentially required subjects to ignore the celebrity pictured in the photo and inhibit themselves from responding to the pictured celebrity, while attending to the name that appeared in text). The researchers found that carrying a working memory load interfered with the ability to inhibit a competing response – or, in the author's terms, reduced the availability of working memory for maintaining priorities that guide visual selective attention. This, in turn, led to greater intrusion of irrelevant distractors.

Figure 8-7. Example Study of Early Working Memory Research

However, such an effect (of cognitive loading on supervisory attention) may also create vulnerability for drivers such that stimuli which should be ignored cannot be (e.g., events outside the vehicle that are irrelevant to safe driving, conversation, or activities of passengers within the vehicle). These types of effects from cognitive task loading were not examined in this study, but should perhaps be followed up in future work.

Findings such as those related to the auditory-vocal tasks in this study underscore the importance of working memory on concurrent task performance, including that done while driving. They confirm the need for some explicit treatment of these issues (both working memory and supervisory attentional functions) in a theoretical framework for driver workload and distraction.

8.4.1.5 Supervisory Attention

In his book, *Understanding Driving*, John Groeger (2000), provides an excellent overview of work that is specifically relevant to the driving task, and, in particular, to the issue of multitasking while driving. Groeger (2000) suggests that human attention and specifically the functions of supervisory attention play a central role in multitasking. They may, in fact, play a central role in understanding the issue of distraction and workload interference while driving and may be important to include in any theoretical framework used to predict which concurrently performed tasks may result in performance decrements.

Groeger (2000) suggests that during driving, different schemata compete for the control of thought and behavior. A schema in this context can be thought of as a routine mental program for control of highly practiced skills. Examples of some schemata that compete for driver attention include: Gear-Changing, Speed-Control, and perhaps some elements of object-and-event detection. The attentional system determines which of the schema that are active and/or vying for attention will in fact get attention, when they get attention, and for how long.

Several functions of the attentional system are characterized by Groeger (2000), and based on work by Stuss, Shallice, Alexander and Picton (1995), Shallice (1982), and others. Many of these functions are now known to involve the pre-frontal cortex (PFC) of the brain, and the areas associated with them are mentioned below each item in the list (from Groeger, 2000, p. 59):

- Setting attentional allocation to a goal (e.g., drive to a specific destination)
  (dorsolateral prefrontal cortex)
- Sustaining preparedness (vigilance to enable response to relatively rare events e.g., unexpected hazards)
  (right lateral mid-frontal regions of the brain, possibly activation/inhibition of target)
- Maximizing activation of current schema (and preparing for upcoming action) (e.g., concentration to manage triggering and coordination of lower-level schemas, e.g., schema for "exiting freeway" and "turning left onto Van Dyke Road") this also includes generating updated expectations which may be associated with "Situation Awareness"
  (anterior cingulated with reciprocal connections to dorsolateral frontal cortex, or circuit comprising connecting midline thalamo, cingulated, and supplementary motor areas)
- Suppressing associated irrelevant schemata (e.g., inhibiting the schema for "overtaking a car" near intended freeway exit)
  (bilateral orbitofrontal areas)
- Sharing across schemata (e.g., listen to radio and drive)
  (orbitofrontal and anterior cingulated regions)
- Switching between schemata (e.g., between lanekeeping and cell-phone dialing)
  (dorsolateral frontal regions of either hemisphere, also more diffuse areas)

As an additional example, D'Esposito et al. (1995) examined patterns of brain activation while participants were performing a spatial task (mental rotation) and a verbal task (semantic verification) concurrently. The key finding of the study was that, when the two tasks are performed simultaneously, the dorsolateral PFC was activated (as well as parts of the brain regions implicated in the performance of the spatial and verbal tasks), even though neither of the tasks activated that particular area (the PFC) by itself. This result suggested that some sort of executive control processes, in which the PFC seemed to play an important role, are implicated in coordination of multiple tasks simultaneously (Miyake and Shah, 1999, p. 464). In short, the role of the pre-frontal cortex in phenomena related to supervisory attention appears consistent with the Wickens concept of "structural" resources.

The functions identified above by Groeger (2000) also bear similarities to concepts described by Posner and Peterson (1990). Their work characterizes the major functions of attention as: (1) orienting to sensory stimuli, (2) engaging in executive control, and (3) maintaining an alert state. Also worth noting here, are the notions of Norman and Shallice (1980), who postulate two basic control mechanisms through which supervisory attentional functions operate:

- Contention scheduling (a bottom-up, data-driven process through which sensory input activates processes, concepts, and goals and "schedules" them to receive attention)
- Supervisory attention which involves:
  - Conscious thoughts about internal states
  - Prioritization of action (this prioritization could be different from that emerging from the contention scheduler and is not necessarily conscious, though it can be).

The significance of this work for the DWM project is that some of the effects may reflect interference by task loading with supervisory attentional functions, a critical element for maintaining successful performance on multiple tasks that are concurrently performed. This finding is one that deserves much more attention in its own right. (Other effects, however, reflect interference by task loading with perceptual functions and psychomotor functions, also critical elements for maintaining successful performance on the road and in traffic). To the extent that the functions of supervisory attention are interfered with by task loads on working memory/central attentional resources, multiple aspects of driving performance would be simultaneously affected (e.g., glance behavior, event detection, speed control). To the extent that perceptual processes are interfered with by task loads on input modalities, or output modalities, multiple aspects of driving performance would also be simultaneously affected, but in different ways. The distinct patterns, resulting from distinct task loadings of the two task types used in this study (auditory-vocal and visual-manual) are consistent with a load theory of attention such as the one Lavie (2005) has put forward, but not by any means definitive in this regard, since other root causes could also give rise to simultaneous effects on multiple dimensions. For this reason, further experimental work in this area is recommended.

8.5 Recommended Toolkit for Use During Product Development

The results of this project make a contribution not only toward the conceptualization and theoretical underpinnings of driver workload and distraction, but also toward the development of practical tools that can be used during product development by vehicle manufacturers or other organizations that may need to evaluate products.

It should be pointed out that a toolkit such as the one described below could be applied within a process of iterative test and re-design, such that a system may be evaluated with as many as four or more independent tests (on separate samples of test participants), before completing its development. Typically, the system would have undergone re-design after each test phase. Note that the sample size across all tests, each of which could be thought of as an opportunity to replicate effects observed in the prior test(s), will be much larger than sample size for each individual test application. Finally, note that this toolkit is designed to provide engineering input to the product development process at all of its stages, from pre-prototype, through bench prototype, to mockup and pre-production (fully drivable) prototypes, as shown in the Figure 8-8. Therefore, one or more tools are recommended for use at each phase of development.

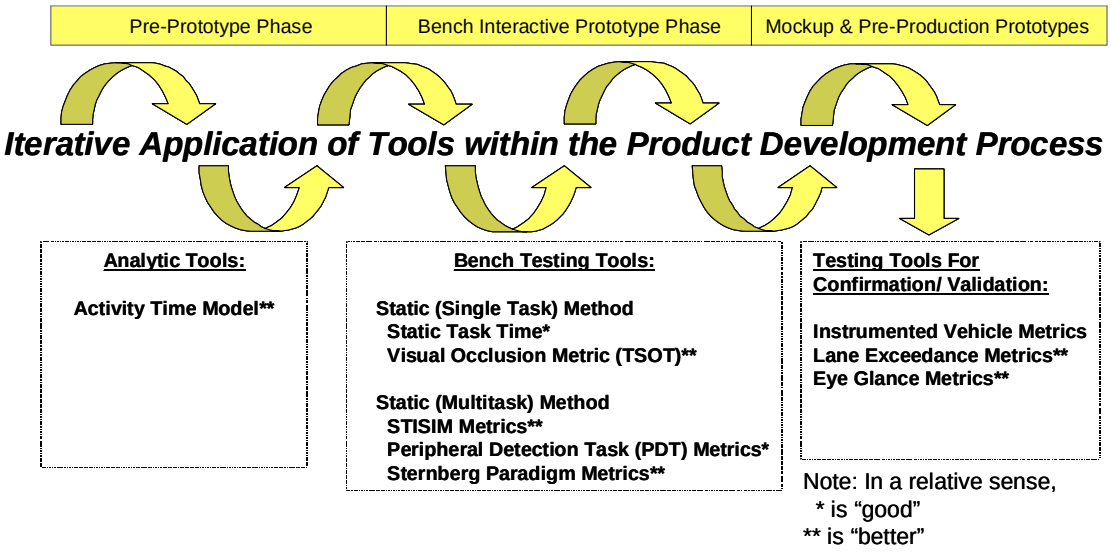


Figure 8-8. Driver Workload Metrics Toolkit

The work done during this project builds upon and extends the work begun by Angell, Young, Hankey, and Dingus (2002) to evaluate alternative methods for assessing driver workload in the early development of advanced in-vehicle systems. The tools recommended below have been selected using three criteria established in this project: (1) repeatability, (2) surrogate validity to predict driving performance on the track and/or road, and (3) to be able to discriminate higher from lower-workload tasks (as defined by prior prediction) within the task type to which they apply. The findings of the DWM research, relative to these criteria, are summarized in Table 8-1 for visual-manual tasks and Table 8-2 for auditory-vocal and mixed-mode tasks. In the tables below, a color coding of green meant it was a good tool that met all criteria, and a coding of yellow meant that tool was recommended but with caution as it was not explicitly evaluated, or was associated with issues that need to be resolved.

Table 8-1. Basis for Toolkit Recommendations for Visual-Manual Tasks

VISUAL-MANUAL TASKS Summary Of Surrogate Metrics Relative To Criteria For Toolkit Selection					
	Repeatable?	**Predictive Validity?**	**Discriminate HIGH from LOW Workload Tasks within Type?**	**Type Of Significant Effect On Driving Performance Effect Addressed**	**Recommended For Toolkit?**
Tool		Visual-Manual	Visual-Manual	(See key at bottom)	
Analytic Modeling Tool					
Total Activity Time	See Text	YES	Not Evaluated	L, S, G	YES
					(discrim. not eval.)
Subjective Rating Tools					
OWL Rating	YES	YES	YES	Subj	YES
Multitasking Difficulty	YES	YES	YES	Subj	YES
Static Time Tool					
Static Time	YES	YES	YES		YES
Visual Occlusion Tool					
TSOT	YES	YES	YES	L, S, G	YES
R-Metric	YES	NO	NO		
Simulated Car Following					
STISIM Duration	YES	YES	YES	L, S, G	YES
STISIM SDLP	YES	YES	NO		
STISIM % Lanex Trials	YES	NO	YES		
STISIM SpeedDiff	YES	YES	YES	L, S	YES
PDT Alone					
PDTA Miss Rate	YES	NO	NO	L	
PDTAMean RT	YES	Not Assessed	NO		
PDT with Simulated Car Following					
PDTS Miss Rate	YES	YES	NO		
PDTS Mean RT	YES	Not Assessed	NO		
Sternberg Assessment Tool					
Strn Pct Miss Rate	YES	YES	NO	E	Yes
Strn Pct AllError	YES	YES	NO	E	Discriminability
StrnCombDecr	YES	YES	NO	E	paradox must be
StrnMeanRTAll	YES	NO	NO		addressed when
StrnMeanRTCorr	YES	NO	NO		used (see text)

Key:
E - Event Detection
G - Glance Behavior
L - Lanekeeping & Lateral Control
S - Speedkeeping & Longitudinal Cnrl.

Subj - Subjective & Experiential

Table 8-2. Basis for Toolkit Recommendations for Auditory-Vocal and Mixed-Mode Tasks

AUDITORY-VOCAL & MIXED MODE TASKS Summary Of Surrogate Metrics Relative To Criteria For Toolkit Selection					
	Repeatable?	**Predictive Validity?**	**Discriminate HIGH from LOW Workload Tasks within Type?**	**Type Of Significant Effect On Driving Performance Effect Addressed**	**Recommended For Toolkit?**
Tool		Auditory-Vocal	Auditory-Vocal	(See key at bottom)	
		(also Mixed-Mode)	(also Mixed-Mode)		
Analytic Modeling Tool					
Total Activity Time	See Text	YES	Not Evaluated	L, S, G, E	YES
					(discrim. not eval.)
Subjective Rating Tools					
OWL Rating	YES	NO	NO	Subj	NO
Multitasking Difficulty	YES	NO	NO	Subj	NO
Static Time Tool					
Static Time	YES	Not Applicable	Not Applicable	NA	NO
Visual Occlusion Tool					
TSOT	YES	Not Applicable	Not Applicable	NA	NO
R-Metric	YES	Not Applicable	Not Applicable	NA	
Simulated Car Following					Possibly, see text
STISIM Duration	YES	Not Applicable	NO	S, L	
STISIM SDLP	YES	NO	NO		
STISIM % Lanex Trials	YES	NO	NO		
STISIM SpeedDiff	YES	YES	NO	S, L	
PDT Alone					
PDTA Miss Rate	YES	YES	YES	E	YES
PDTAMean RT	YES	NO	YES		
PDT with Simulated Car Following					
PDTS Miss Rate	YES	YES	YES	E	YES
PDTS Mean RT	YES	NO	NO		
Sternberg Assessment Tool					
Strn Pct Miss Rate	YES	YES	NO		
Strn Pct AllError	YES	YES	YES	E, G	YES
StrnCombDecr	YES	YES	YES	E, G	YES
StrnMeanRTAll	YES	YES	NO		
StrnMeanRTCorr	YES	YES	NO		

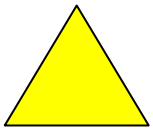
Key:
E - Event Detection
G - Glance Behavior
L - Lanekeeping & Lateral Control
S - Speedkeeping & Longitudinal Cnrl.

Subj - Subjective & Experiential
NA - Not Applicable

The rightmost column of Table 8-1 and Table 8-2 identify the tools that may be appropriate for organizations doing evaluations of advanced in-vehicle systems to include in their toolkits for assessment (at least insofar as the results of this study and the criteria used here would indicate).

In describing the recommended toolkit, the reader will find descriptions of each tool, along with an identification of the recommended metrics produced by the tool, the types of tasks to which they may be applied, and the uses to which they may be put (for example, for identifying high-workload tasks that may need design improvements, or for making early predictions of what driving performance effects might result if the system were used on the road). It should, of

course, be understood that any prediction from a surrogate method should be considered along with other information sources for affirmation. These other sources of information include comparable systems analysis, previous evaluations, real-world experience with the product under evaluation, test results obtained with different methods, and consensus engineering judgment. None of these information sources is perfect. But greater confidence can be placed in several imperfect sources that all point toward the same conclusion. This will provide a context of sound decision processes to avoid errors in the interpretation and application of surrogate measures. . The importance of this cannot be emphasized enough. The relationships between surrogate measures and driving performance measures are imperfect, and the toolkit must be used with this caution in mind.



Caution: Surrogate Metrics are imperfect indicators of driving performance and task effects. Use them with caution; interpret any findings with great care.

In addition, there are several difficult issues that emerged in the selection of tools for the toolkit. These will be discussed after it is described. For now, suffice it to say that the evaluation methods for assessing task workload and its various effects on driving performance are at an early stage of evolution, and will continue to develop and change as the science behind them deepens.

8.5.1 Tools for Use during Pre-Prototype Product Development

8.5.1.1 Activity Time Model

To enable prediction of driving performance effects early in product development, an analytic model is recommended. In particular, in this study, the Activity Time Model proved the most useful of the models evaluated in the study. It requires as input data that can typically be obtained very early during the development process, thus making it amenable to application even before bench prototypes may be available.

In the modeling tool recommended for use here, modelers analyze tasks and identify cognitive and physical activity steps required to perform a given task step. Then task steps are assigned activity times, based simple models and typical values of physical, sensory-motor, and cognitive processes taken from the literature. Task steps are also assigned resource requirements based on Modified MRT (as implemented under this project). It is recommended that more than one analyst model each task to identify differences in task execution strategy and improve inter-rater reliability issues addressed in Chapter 6. Subsequently, results from all modelers should be combined to identify different user strategies for task performance, and/or to improve repeatability of task modeling. Further development of analytic models is warranted, especially as they are applied and used.

Recommended Metrics:	Total Activity Time (per task)
Applicable To:	Visual-manual tasks, auditory-vocal tasks (and mixed-mode)
Useful For:	Identifying tasks during the pre-prototype phase that may need further development before testing begins.

	Ability to discriminate high from low-workload tasks within each type was not formally examined in the same way as for other measures.
Has Predictive Validity For:	Visual-manual: Task Duration, SDLP, Speed Difference, Percent Lane Exceedance Cross Trials, Number and Duration of glances to Road, Not Road, SA, MR, TR
	Auditory-vocal: Road LVD and CHMSL Miss Rates, Track SDLP and Speed Difference, Mean Road Single-Glance Duration

8.5.2 Tools for Use With Early Prototypes

There are two types of tools or methods described below that may be used with early prototypes: single task and multitask. The single task methods allow the test participant to devote full attention to the use of the in-vehicle device. The multitask methods require the test participant to divide attention between the use of the in-vehicle device and at least one other task (though some of them are simple and easy to implement).

8.5.2.1 Static Single-Task Methods

The methods below may be administered with only a bench top representation of a system and the tools needed for the assessment. In these methods, a participant is allowed to devote full attention to the in-vehicle task(s) of interest (and is not asked to divide attention between a driving-like task and the in-vehicle task).

Static Task Time Measurement Tool

This tool consists of a laboratory setup for collecting task completion times. It implements a method (e.g., similar to that in SAE J2364 (2004)) in which a participant performs a task without any other concurrent task or interruption. This method has been validated for use only with visual-manual tasks. The method generates a metric of task completion time, and requires careful and consistent definition of task start and end states. For the DWM study, timing of the task began when the experimenter said the word “now” in the templated task request which included the words, “Please begin now.” (Example: “Your task is to tune the radio to 104.3. Please begin now.”) The task time ended when the participant said “Done” upon completing the task (in this example, when they reached the frequency 104.3). This method may be applied using a stopwatch, but may also be applied using computer-based software, such as that developed and used in this study.

Recommended Metric:	Task Completion Time
Applicable To:	Visual-manual tasks only (applicability to duration-intrinsic auditory-vocal or mixed-mode tasks is unknown)
Useful For:	Identifying high-workload tasks in the laboratory (that may need further design/test cycles)

Has Predictive Validity For:	Number of Task-Related Glances (track)
	Total Eyes Off-Road Time (not road) (track)
	Task Duration (while driving) (track)
	Speed Difference (track)

Visual Occlusion

This tool consists of special-purpose equipment and software that permit a test participant to see the task they are performing only in brief glimpses until the task is finished. The method is one in which a participant performs a task wearing special goggles equipped with a computer-controlled shutter. In this study, the shutter was programmed to open and close repeatedly throughout a task using a cycle consisting of 1.5 seconds open period and 2.0 seconds closed period. Other closed interval times ranging from 1.0 seconds to 2.0 seconds could be considered to be compatible with on-going standards development efforts for visual occlusion in other organizations. This method can be meaningfully applied only to visual-manual tasks. This method generates a number of metrics, but the one validated by analyses to date is TSOT. For a description of the equipment used to implement this method, see Chapter 5 and Appendix D.

Recommended Metric:	Total Shutter Open Time (TSOT)
Applicable To:	Visual-manual tasks only
Useful For:	Identifying high-workload tasks in the laboratory (that may need further design/test cycles)
Has Predictive Validity For:	Task Duration (road and track), SDLP (road and track), Percent Lane Exceedance Cross Trials (road), Speed Difference (road and track), Number of Task-Related Glances (road and track), and Mean Task-Related Total Glance Duration (road and track)

8.5.2.2 Static (Multi-task) Methods

These methods may also be administered with a bench top representation of the system, but require that the participant divide attention between the test method and the in-vehicle task of interest.

Peripheral Detection Task Alone

This tool allows a high-intensity spot of red light to be briefly projected (producing a stimulus duration of 1 second) onto a blank screen in front of the participant during task performance. The location of the light on the screen was 10 degrees left of the centerline of the driver's seat. The participant can activate a pushbutton in response to detecting the light. The participant is asked to perform each in-vehicle task of interest while also monitoring for the appearance of these lights and responding to them as rapidly as possible when they are detected. The lights are presented at inter-stimulus intervals taken from a uniform distribution ranging from two to 10 seconds. The metric recommended for use based on this study is Percent Missed Events. See Chapter 5 and Appendix D for

further details on the equipment and software configurations that can be used to implement this method.

Recommended Metric:	Percent Missed Events
Applicable To:	Auditory-vocal tasks (and visual-manual tasks pending an adjustment for the duration paradox to aid in interpretation)
Useful For:	Identifying tasks in the laboratory that may affect event detection and that may benefit from further design/test cycles.
Has Predictive Validity For:	Auditory-vocal tasks: Percent CHMSL Miss Rate (road only)

Simulated Car Following With Peripheral Detection Task

The peripheral detection task described above can also be used in conjunction with a simulated car-following scenario that is implemented using a fixed-base, part-task driving simulator (e.g., the STISIM *Drive*™ from Systems Technology, Inc. that was used in this study). When used in this way, the PDT light appears on the driving scene in the same location as it would have appeared had the screen been blank (as above). PDT stimulus presentation was coordinated with the STISIM driving simulation through special-purpose software and hardware needed to link two PCs (the STISIM ran on its own computer). See Chapter 5 and Appendix D for details of the STISIM *Drive*™ simulator used in this project. This method produces lanekeeping and speed maintenance metrics in addition to PDT measures. The metrics recommended for use is listed below, along with the driving performance measures that it can be used to predict. (Please use the detailed results of Chapter 5 to guide application.)

Recommended Metrics:	**Predictive Validity**
Median STISIM Task Duration	Task Duration, SDLP, Speed Difference, Percent Lane Exceedance Cross (road only), Task-Related Eyeglance Counts, Total Eyes Off Road (not road) Time
Median STISIM SDLP	SDLP, Speed Difference, Task Duration, Percent Lane Exceedance Cross
Median STISIM Speed Difference	Speed Difference, SDLP, Task Duration
PDTS Percent Miss Rates	CHMSL and FVTS Miss Rates (Road)
Applicable To:	Visual-manual tasks (all metrics, except PDTS) Special Note on PDT-STISIM Percent Miss Rate: This may possibly apply to visual-manual tasks (but there is a paradox concerning discriminability (see section 8.2.2.4 for a full

	explanation) this method was not intended for use with short visual-manual tasks)
	Auditory-vocal tasks (only one metric applies): PDTS Percent Miss Rate (correlates with Percent CHMSL Miss Rate on Road) However, STISIM Task Duration and STISIM Speed Difference did have predictive validity but not discriminability for the set of auditory-vocal tasks evaluated in this study.
Useful For:	Identifying high-workload tasks (that may need further design/test cycles)

Sternberg Memory Task

This method involved remembering and recalling road signs when probes appeared during performance of an in-vehicle task. In many respects, it is similar to the Peripheral Detection Task Alone method, except that road signs instead of lights are presented, and the participant must carry a memory load throughout the duration of a task. While a participant performs a task, a road sign is briefly presented on a display. The participant is asked to press one pushbutton if the displayed sign is from a set of signs memorized prior to the start of the task, or a second pushbutton if not. One version of this method used route junction signs while another version used route number signs. The version using route junction signs was more difficult, leading to longer response times and more errors (and more interactions with age) than the version using route number signs. However, the results upon which our recommendations are based pooled data from both versions, so both versions should be used in the toolkit as applied in the study done here. See Appendix R for details. The metrics generated by this method which are recommended for use are:

Recommended Metrics:

Sternberg Metrics	**Predictive Validity**
Sternberg Percent Missed Detections	Percent CHMSL and FVTS Miss Rate (road) and various eyeglance metrics
Sternberg Percent All Errors	Percent CHMSL and FVTS Miss Rate (road) and various eyeglance metrics
Sternberg Median All RT	Percent LVD Miss Rate (road for auditory-vocal tasks) and various eyeglance metrics
Sternberg Combined Decrement Score	Percent CHMSL and FVTS Miss Rate on-road and various eyeglance metrics
Applicable To:	Auditory-vocal tasks (and mixed-mode tasks)
	Possibly to visual-manual tasks (but there is a paradox concerning discriminability; please see Issues in this section for a full explanation; this method was not intended for use with short visual-manual tasks)

Useful For: Identifying tasks that have intrusion on event detection and may need another re-design/test cycle.

8.5.3 Tools for Use With Drivable Vehicles (for Confirmation and Validation)

Usually, if it is necessary to carry out confirmation or validation testing with a full-scale drivable prototype vehicle, it will take place first on a closed-course test track, and perhaps subsequently, and only in certain circumstances, on open roads. It is therefore important to note that test track and road effects do differ somewhat, perhaps because of the effects of traffic density differences, among others. The tools needed for this work are considerably more complex than those in the remaining parts of the toolkit. The interested reader is referred to Appendix C for descriptions of the equipment and software necessary for instrumenting vehicles properly for this kind of work.

Methodologically, there are four important things to note. First, it is important to note that in evaluations of workload demands imposed by systems on drivers, this research identified the need for each task to be run with at least one trial, and preferably two trials, free from any events-to-be-detected, so that pure measures of glance behavior to the device, road, and mirrors can be obtained. Second, it is also desirable that each task be run with at least one trial, and preferably two, that include events-to-be-detected so that measurements of responsiveness to events can also be obtained. Third, it is desirable to include a Just Drive task for use as one benchmark against which to compare other tasks. Finally, tests must be administered using a large enough sample of participants to ensure adequate statistical power to guard against errors in decision making. Determining statistical power requires a knowledge of both the size of the effect to be detected and variability of the metrics used. Readers interested in more information on statistical power calculations should consult a text on the subject such as Cohen (1988).

Recommended Metrics: It is recommended that metrics be obtained in each of the following four categories to comprehensively assess driving performance.

- Event Detection
- Eyeglance
- Lateral Control[1]
- Longitudinal Control

Applicable To:

- Visual-manual tasks
- Auditory-vocal tasks
- Mixed-mode tasks

Useful For: If needed, to confirm that tasks have been successfully designed and tested to meet workload targets through prior test-and-redesign cycles.

[1] Note: Use caution in interpreting lane exceedance data from test track evaluations. In the DWM study, lane exceedance from the track did not have predictive validity for the road. This may have been due to any number of factors that differed in our study between the track and road venues, including differences in tasks tested, traffic density, and safety observer precautions on the public roadway. Nonetheless, when possible, validation under on-road conditions is always ideal.

8.5.4 Subjective Rating Tools

Some vehicle manufacturers may be interested in their customers' subjective experience of workload, and some evaluators may feel a driver's subjective experience may influence the strategies that he/she may utilize in deciding when to initiate a task, how to perform a task, how to adjust other tasks to accommodate a newly-initiated one, or whether to suspend a task. For these purposes, it may be useful to include one or more subjective rating tools in the assessment toolkit. Two such rating tools proved useful in this work for visual-manual tasks, the Operator Workload Level (OWL) rating tool, and the Multitasking Difficulty rating tool. Examples of these tools as used in this study are provided in Appendix I.

For evaluators who may choose the OWL rating tool, it is important to realize that the range of workload among the task set being evaluated may influence task ratings. If so, a particular task will receive a different rating when rated in the context of different task sets (if one set has a wide range of workload and the other a narrow range of workload). In a production environment, this variability in ratings for a given task on a given device can lead to misunderstanding and misinterpretation. To prevent this from happening, care should be taken to ensure that a common high- and low-workload task are inserted into all task sets that are likely to be compared.

For the Multitasking Difficulty rating tool, instructions to participants are important. If properly instructed, the rating technique is supposed to yield values that are stable across task sets (regardless of the range of workload included in the task set). This technique is usually more robust to task-set differences though it is not immune from them either (Gescheider, 1997). However, the application of magnitude estimation techniques to workload scaling in this study is believed to be the first attempt, so there has not yet been an opportunity to examine this question of task-set differences empirically in the context of driver workload.

Methodologically, it is usually optimal to obtain ratings of subjective workload experience immediately after task performance (if time permits), and to obtain more than one observation (averaging ratings collected after each of two trials, for instance).

Recommended Metrics:	OWL Rating
	Magnitude Estimate
Applicable To:	Visual-manual, auditory-vocal, and mixed-mode tasks
Useful For:	If needed, to identify tasks for which re-design/test may be desirable.
Has Predictive Validity For:	Task Duration (track and road)
	SDLP (track)
	Speed Difference (track and road)
	Number of task-related glances (high on-road, less on track)
	Mean Duration of Task-Related Glances (road)
	Total Eyes Off-Road (not road) (road and track)
	Percent Lane Exceedance Cross Trials (high on-road, not on track)

8.5.5 Star Charts: A Decision Aid for Tasks during Product Development

One tool that is particularly needed by OEMs is a method for graphically depicting all tasks in a way that allows direct comparison of their multidimensional effects on driving, to support decisions that must be made regarding a task's compliance with internal company requirements for managing driver workload, as well as external guidelines and standards. Such a method should ideally also support the need for re-design of a task, and assist (if possible) in diagnosing the area in which re-design may be needed.

A type of tool that can be useful in this regard is that represented by the star charts presented in Chapter 3, Figure 3-58 and also in Figures 8-4, 8-5, and in Appendix S. As explained elsewhere, these star charts show the scores of a task on each of several performance metrics. Each performance metric is illustrated on a radial in the star chart, and its scale is normalized. This allows straightforward comparisons of tasks to determine on which radials scores are high, to determine whether overall area of a task (representing the extent of its effects) is larger than some desired criterion. Most importantly, star charts allow a task's effects on multiple dimensions to be simultaneously illustrated in a way that is easily grasped and compared with the effects of other tasks on the same set of dimensions.

To construct a star chart, the task-level scores on a metric (e.g., median static task time) need to be converted to z-scores, using standard statistical formulas (and the mean and standard deviation for the sample of task scores on that metric). The resulting set of scores will have a mean of zero and a standard deviation of 1.0. This process must be repeated for each metric to be displayed as a radial on the star chart.

Charts constructed in this way can be compared to support decision-making. Specifically, the shape and sizes of the star for each task of interest reveals how it compares with other tasks in the set against which it was standardized. If a score falls above zero on a radial, this indicates the task is above the average on that measure, and may suggest that the task could benefit from design improvements.

Of course, it will always be important to include a star chart for just driving as a baseline comparison chart whenever diagnoses and decisions of this sort are being made.

In addition, if an organization has established criteria, or cut-off points, which reflect levels on each metric above which scores are deemed to identify tasks in need of re-design, then those cut-off points can be plotted and color-coded as a star or annulus for use in comparing a task to the thresholds it must meet.

8.5.6 Issues with Toolkit Contents, Coverage, and Use

A toolkit should contain a tool for each type of assessment that needs to be undertaken in evaluating the effects of task workload on driving performance. Ideally, the toolkit would consist of a single general-purpose tool that handles most issues. This is not the case in task-related driver workload prediction.

To review, DWM visual-manual and auditory-vocal tasks had effects on event detection and eyeglance behavior that were generally discriminable from just driving. The effects for visual-manual tasks were larger than those for auditory-vocal tasks, but a paradox emerged for event detection in which visual-manual tasks affected event detection, in ways contrary to their expected workload level. Specifically, it was not possible to discriminate between higher-workload and lower-workload visual-manual tasks (as defined by prior prediction) on the dimension of event detection. A task duration effect was discovered that introduced a paradox. Shorter duration tasks were associated with poorer detection performance when prior prediction indicated that shorter tasks should be lower in workload. Therefore, none of the event detection

methods formally met the criterion of discriminability that would allow them to be recommended for the toolkit for visual-manual tasks. Yet the event detection effects were among the most prominent findings in the research, and certainly deemed relevant for assessment if an appropriate methodology can be found. This set of issues created a problem for the toolkit (from the standpoint of completeness of coverage) that is discussed in the next section.

Visual-manual tasks had effects on lanekeeping speedkeeping and task duration that were generally discirminable from both Just Drive and between higher-workload and lower-workload tasks. The fact that difficult tasks led to effects on vehicle control measures underscored the importance of continuing to evaluate for effects of tasks on lateral and longitudinal control. However, for auditory-vocal tasks, no surrogate proved adequate, thus creating an issue of toolkit completeness in the categories of lateral and longitudinal performance measurement.

The visual-manual tasks and auditory-vocal tasks therefore had contrasting measurement profiles. Visual-manual tasks were associated with lateral control, longitudinal control, and task duration effects that were repeatable and were predicted from laboratory surrogate measures. Higher and lower-workload visual-manual task effects were also discriminable on these measures. (Of course, visual-manual task durations were shorter than Just Drive's fixed two minute duration). On the other hand, Object and Event Detection measures were not readily interpretable due to a paradox in which shorter duration tasks had worse OED performance than longer tasks. Interestingly, this effect was also present for a short-duration auditory-vocal task. This paradox appears to be more by task duration than by task type. This same effect was found for the shorter auditory-vocal task of Book-on-Tape Summarize. The task duration paradox for event detection appears to be methodological in nature and so may be addressed by methodological adjustments.

Now consider the auditory-vocal tasks. Higher- and lower-workload auditory-vocal tasks were generally not well discriminated by vehicle control measures in any venue. The absence of effects for auditory-vocal tasks in vehicle control measures is not necessarily methodological. It may reflect a legitimate minimal impact of the DWM auditory-vocal tasks on vehicle control. As discussed earlier, automaticity does not appear in driver gear-changing operations. But other aspects of driving performance may be more reflexive. For example, a moment's reflection makes clear that continuous lanekeeping is probably one of the most heavily practiced of manual skills for most people (Hancock, 2005). Millions of people drive for hours a day, every day, year after year. Continuous lanekeeping is ever-present while driving and may be categorically different than intermittent gear changes. Brown's theory of selective withdrawal of attention (Brown, 1984) has been mentioned several times in this report. It provides a plausible explanation for the absence of lanekeeping effects for the DWM auditory-vocal tasks. However, vehicle control effects might arise if auditory-vocal tasks are more demanding, e.g., have emotional content, involve extended periods of debate-like conversation, or involve complex device operations. It is also important to remember that all DWM tests with simulated or real driving were on relatively straight or gently curving roads. (The curves on the test track were one exception, but a highly predictable one). It is therefore possible if not also plausible to expect that a more demanding (e.g., serpentine) course might reveal auditory-vocal task effects on lanekeeping or speed control or both. And it is possible that auditory-vocal tasks of different durations might show vehicle control effects. This is suggested by the Book-on-Tape Summary task relative to the longer duration auditory-vocal tasks. For these reasons, vehicle control effects (or estimates thereof) should be included in workload predictions for auditory-vocal tasks.

In the object and event detection methodology used here, events were scheduled to occur at some point within every task, no matter how short. This was done so that measures could be taken of responsiveness to events during task performance (with the notion that it may be valuable to know how responsive a driver may be in the circumstance that an event might occur during such a task). However, visual-manual tasks, which involve removing the eyes from the forward road

scene at least some of the time during a task, all affected event detection in this methodology. The interactions between task length, frequency of engagement, and co-occurrence of events that may exist in "real" driving were not reflected in this research. This research reflects only the effects that the visual demand of a task may have on event detection when an event occurs during a task. They should not be taken to suggest anything at all about the risk associated with the tasks, since risk is a concept that involves many other variables besides the demand on the driver, and its interference with driving (not the least of which is the probability with which an event may co-occur with a task of a given length and frequency). It may even be that for tasks which require visual-manual interactions, the best way to mitigate effects on event detection would be to design such interactions to be short and infrequently executed (based on the notion that it is much less likely that an event will co-occur with such a task than with a longer, frequently done task). These complicated issues of risk management were outside of the scope of this project, and are therefore left to others to deliberate and resolve.

From the perspective of the toolkit, what is needed is a means of applying the event detection methodology to visual-manual tasks (or perhaps all kinds of tasks that tend to be shorter in duration) that might allow discrimination across tasks (or minimally from some criterion level of event detection that an organization might deem to be acceptable). Several means to address the task duration paradox were presented earlier in this chapter. All of these proposed methods have limitations that may limit their usefulness. There may be ways to set such a criterion in the use of event detection measures that would enable this method to be used. Or there may be other ways to address the issues (for example, applying the method only to longer, new visual-manual tasks for which it was really intended in advanced information systems). Again, these issues of application are left to the organizations that may use the toolkit to consider and resolve for themselves. However, if it should be determined that none of the methods having predictive validity for event detection performance during visual-manual tasks can be used because no suitable solution to these issues can be found, then there will be a serious issue of incompleteness with the toolkit, because assessing the effects of visual-manual tasks on event detection appears from this research to be important. We would therefore identify this area as one that could benefit from further work.

8.5.6.1 Driver Eyeglance Assessment

No direct measures of glance behavior in the laboratory were recommended for the toolkit, based on the assumption that they may require too much labor, time, and cost to be practical. However, for an organization that has the resources, glance measures acquired in the laboratory may well be useful (though they would need to be validated against glance measures gathered on the road).

8.5.6.2 Lateral Control Assessment

No surrogate measures proved to have either predictive validity or discriminability for auditory-vocal tasks in the area of lanekeeping or lateral control assessment. This is problematic, insofar as the toolkit is incomplete without such measurements. As indicated above, the absence of lanekeeping effects for the DWM auditory-vocal tasks may be may be real. As suggested earlier, it is possible that the restricted range of task lengths for auditory-vocal tasks in this study contributed to the inability to discriminate among these tasks on this dimension. Contrary to the auditory-vocal task results, lateral control measures were useful for visual-manual tasks. Overall, it is suggested that lateral control measures be included in a toolkit for auditory-vocal and mixed-mode tasks as well as for visual-manual tasks. This applies to laboratory simulators, test track, or on-road evaluations in order to obtain longitudinal control as well as lateral control assessments. Perhaps other simulators not examined here will prove useful for predicting and discriminating auditory-vocal and mixed-mode tasks in the category of lateral control performance measures.

Finally, CAMP DWM procedures may be fully capable of capturing the effects of more taxing auditory-vocal tasks than those used in the study.

8.5.6.3 Longitudinal Control Assessment

No surrogate measures proved to be able to discriminate higher from lower-workload auditory-vocal tasks on the measure of Speed Difference (though the STISIM measures of Speed Difference did have predictive validity). The Speed Difference measure was able to discriminate among visual-manual tasks. With this category of measure, as with the lateral control category, it is possible that the restricted range of task lengths for auditory-vocal tasks in this study contributed to the inability to discriminate among them on this dimension. However, the fact that there is no tool to offer for auditory-vocal and mixed-mode tasks in this category of measurement is also problematic, insofar as the toolkit is at this time incomplete in this area as well. Comprehensive assessment suggests that in the near term it may be desirable to include longitudinal control measures like Speed Difference in a toolkit for auditory-vocal, mixed-mode and visual-manual tasks, assessed in simulator, test track, or road evaluations. Perhaps other simulators or other surrogate tools not examined here will prove useful for predicting and discriminating auditory-vocal and mixed-mode tasks in the category of longitudinal control performance measures).

8.5.6.4 Fair and Even-handed Assessments between Task Types

In evaluating tasks of different types, visual-manual versus auditory-vocal versus mixed-mode tasks, it will be important for evaluators and decision-makers alike to create a set of comparisons that are fair and even-handed with respect to the whole set of tasks (especially since in some systems, all types of tasks occur). It would not make sense, for example, to evaluate auditory-vocal tasks for event detection effects (because the surrogates discriminate) but not do so for visual-manual tasks (when the effects of visual-manual tasks on event detection are much larger than those of auditory-vocal tasks). Similarly, it would not make sense to evaluate visual-manual tasks relative to Speed Difference or SDLP or Lane Exceeds, but not do so for auditory-vocal tasks that may impose high workload (simply because the surrogate tool happened to work for the visual-manual tasks, but not for auditory-vocal tasks of moderate difficulty). It is therefore extremely important that the issues sketched forth above be resolved by individual organizations in a way that places them in the context of the full set of task decisions that must be made. Otherwise, it is quite possible that erroneous decisions about tasks and systems may result.

8.5.7 Summary of Toolkit

To review, the toolkit that emerged from this work is one that is equipped to provide iterative support to a product development process throughout all of its phases, from pre-prototype through to production-readiness. In addition, across the tools in the toolkit, an effort was made to provide tools which cover all four categories of driving performance measurement for each task type– event detection, eyeglance behavior, lateral control, and longitudinal control.

However, several difficulties emerged with respect to this effort. First, the task types have different effects relative to just driving. Not all categories of effect were discriminable from just driving for both types of tasks. Discriminability between higher-workload and lower-workload tasks within a category was sometimes low even if a measure could distinguish tasks from Just Drive. Second, there were significant issues emerging from methodologies and task set constraints that leave some puzzles to be resolved by those researchers who may undertake future work in this field, and/or by practitioners who may more urgently need to find answers to these questions that will suffice in the field. Nonetheless, the toolkit that has been equipped by this project contains methods that have been rigorously evaluated and will prove helpful during the development and evaluation of advanced information systems. It is hoped that it will serve as

strong foundation for further research, while also inviting the development of new and better tools.

8.6 Recommendations for Future Work

A variety of driver distraction issues are worthy of further research. Some of these issues are listed below, in no particular order and without regard to their feasibility:

- **Risk exposure.** When, where, for how long, in what sequence, and how often in-vehicle tasks are typically done. Empirical exposure data would complement driver workload measures to refine estimates of distraction potential.
- **Driver motivation in task completion.** Currently a topic of debate, data are needed on the extent to which drivers will stretch out the completion of a task or get on with it and how this varies with current driving conditions and driver states and traits. Such data would provide an empirical basis to judge the validity of such concepts as interruptibility.
- **Interruption costs.** This issue arises as drivers must switch between the driving task and in-vehicle activities. There is little empirical data (e.g., Monk, Boehm-Davis, and Trafton, 2004) on the costs of task interruption in driving such as forgetting, reorientation time, impact on search processes, frustration, and so forth. This research could lead to countermeasures aimed at specific decrements or driver training to minimize their impacts.
- **What attracts attention in the vehicle?** This question has been investigated to some extent for roadway objects and events (Cole and Hughes, 1988). In-vehicle distraction sources deserve similar scrutiny as inputs to automotive Human-Machine Interface (HMI) design.
- **Lost-in-Thought phenomena.** Tasks may induce distraction but not all distraction is due to tasks. It would be a great breakthrough to understand the time course and indicators of Lost-in-Thought Phenomena as well as the features in the world that draw us back into it.
- **Distraction criteria.** Driver workload measures are very indirect indicators of distraction potential in the field. But practical decisions must be made about in-vehicle design and accessibility by the driver. Research is needed on the alternatives and impact of different decision criteria or thresholds for limiting distraction potential.
- **Distraction effects of voice-recognition systems:** Voice recognition systems to control in-vehicle systems have evolved a great deal over the last decade. Assessment of the distraction potential of voice-recognition systems, which can be designed, would complement the extensive research literature on cell phone conversations, which are difficult to design.

Additional ideas for future experimental work related to needed work on event detection, on the extent to which the experimental results of this study may generalize to naturalistic driving, and to furthering the work on multivariate analyses of the data in exploring the underlying dimensions of driver workload and distraction. These include:

- **Naturalistic driving study of in-vehicle tasks**. Examining performance of in-vehicle tasks under naturalistic driving conditions could provide useful new data. It could help determine whether event-detection effects on eyeglance may

generalize to natural settings, as well as provide information on frequency of engagement, conditions of engagement, etc.

- **Naturalistic driving study of skill acquisition and strategies of use.** Performance may change over time as a driver develops skill with a new system, and strategies for use may form. A study to gather data on this would provide new insight into natural use not currently available from experimental studies.
- **Extending research on event detection and its effect on eyeglance behavior**
  - Exploring effects of event type on task performance variables (such as suspension and/or shedding of a task at event detection, and then resuming a task following event detection, and whether this interacts with task type)
  - Exploring allocations of attention-across the visual field under different task loadings
  - Exploring heightened situation awareness to salient areas following detection of a hazard (event) (e.g., examining whether increased looks to mirror, and increased responsiveness to FVTS, is appropriate in the context of LVD detections)
- **Carrying out proposed Principle Component Analysis and other Multivariate Analysis work**
  - Applying PCA at the task level using road data
  - Applying PCA at the task level to data from the test track venue
  - Apply PCA at the task level to data the from laboratory venue

8.7 Concluding Comments

The DWM project yielded the following insights about driver workload.

- States of driver workload which produced overload or interference with driving performance were manifest not on just one underlying dimension of performance, but on several simultaneously affected ones, confirming that workload-induced distraction is multidimensional in nature.
- There were specific patterns of effects that appeared to be interpretable as characteristic of distraction, but which will require further research to confirm. Visual-manual tasks led to more pronounced intrusion on driving than did auditory-vocal tasks. This intrusion was discriminable from just driving for both types of tasks, but much more subtle for auditory-vocal tasks.
- Different patterns of interference/degradation across the categories of performance were associated with, tasks of different types (auditory-vocal versus visual-manual). However, for both types of tasks, eyeglance measures and event detection measures were key in evaluating extent of intrusion on driving performance.
- Different patterns emerged for individual tasks, which were unique to the demands that each imposed on drivers, indicating that specific structural interference between concurrently performed tasks continues to be an important theoretical construct in understanding what gives rise to performance degradation.

- Event detection affected eyeglance patterns in important ways, a finding that has both theoretical and methodological significance. The implications of this finding included the fact that methods used to study event detection may affect the behavior of interest and the suggestion that when evaluating the visual demand of tasks in an advanced information system or in-vehicle device, it is important that multiple test trials be conducted—some with and some without event detection. The trials used to evaluate the visual demand of a task should not include events to-be-detected in order to obtain clean measurements of glance behavior, free from the influence of co-occurring events.
- A set of surrogate measures was identified for each type of tasks that could be used at each of common product development processes. These surrogate measures were repeatable, meaningful (i.e., were valid for predicting driving performance effects), and enabled discrimination of high from low-workload tasks (within the framework of the study). A toolkit of methods was described that would equip organizations to apply these measures in the evaluation of driver workload during development of advanced in-vehicle systems.

8.8 Chapter References

Alliance of Automotive Manufacturers (AAM). (2003). *Statement of principles, criteria, and verification procedures on driver interactions with advanced in-vehicle information and communications systems* (Version 2.1). Southfield, MI: Author

Angell, L.S., Young, R. A., Hankey, J. M., and Dingus, T. A. (2002). An evaluation of alternative methods for assessing driver workload in the early development of in-vehicle information systems, *SAE Proceedings*, 2002-01-1981, Government/Industry Meeting, May 15, Washington, DC, USA.

Brown, I. D. (1994). Driver fatigue. *Human Factors*, *36*(2), 298-314.

Hancock, Peter. (2001). Personal communication.

Card, S., Moran, T., and Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Erlbaum.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical methods for data analysis*. Belmont, CA: Wadsworth.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (Second Edition).* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, P.K., and Hughes, B. L. (1988). What attracts attention while driving? Ergonomics, 29, 377-391.

de Fockert, J. W., Rees, G., Frith, C. D., and Lavie, N. (2001). The role of working memory in visual selective attention. Science, Vol. 291, 2001, p. 1803-1806.

D'Esposito, M., Detre, J.A., Alsop, D.C., Shin, R.K., Atlas, S., and Grossman, M. (1995). The neural basis of the central executive system of working memory. Nature, 378, 279-281.

Dingus, T.A., Antin, J. F., Hulse, M.C., and Wierwille, W.W. (1989). Attentional demand requirements of an automobile moving-map navigation system. Transportation Research Record, 23A(4), 301-315

Dingus, T., McGehee, D., Hulse, M., Jahns, S., Manakkal, N., Mollenhauer, M., and Fleischman, R. (1995). TraTek evaluation task C3- Camera car study (Report No. FHWA-RD-94-076). Washington, DC: U.S. Department of Transportation, Federal Highway Administration.

Dingus, T., Hulse, M., Jahns, S., Mollenhauer, M., and Fleischman, R., McGehee, D. and Manakkal, N., (1997). Effects of age, system experience, and navigation technique on driving with an Advanced Traveler Information System. Human Factors, 39(2), 177-

Duncan, J., Williams, P., Nimro-Smith, I., and Brown, I. D. (1992). The control of skilled behavior: Learning, intelligence and distraction. In D. E. Meyer and S. Kornblum (Eds.), Attention and performance XIV. Cambridge, MA: MIT Press.

Fancher, P., Ervin, R., Sayer, J., Hagan, M., Bogard, S., Bereket, Z., Mefford, M., and Haugen, J. (1998). Intelligent cruise control field operational test (Report No. DOT HS 808 849). Washington, DC: U.S. Department of Transportation National Highway Traffic Safety Administration.

Fitts, P. M. and Posner, M. I. (1967). Human performance. Belmont, CA: Brooks/Cole Publishing.

Foley, J. P. (2005). Lessons learned from J2364 (Paper No. SAE 2005-01-1847). Warrendale, PA: Society of Automotive Engineers.

Foley, J., Glassco, R., Cohen, D., and Chang, J. (2004). DOT/Mitretek Analysis Of CAMP On-Road Driving Performance Data. Briefing Prepared As Part of CAMP Driver Workload Metrics Proprietary Interim Work Product. January 29, 2004.

Gawron V.J., (2000). Workshop 12: Guide to measuring Workload and Situational Awareness. XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Society, July 29-August 4.

Gescheider, G. A. (1997). *Psychophysics: The fundamentals* (Third edition) (pp. 231-263). Mahwah, NJ: Lawrence Erlbaum.

Gigenrenzer, G., Todd, P. and ABC Research Group (2000*). Simple heuristics that make us smart*. Oxford: Oxford University Press.

Goodman, M.J., Barker, J., and Monk, C. (2005, February*). A Bibliography of Research Related to the Use of Wireless Communications Devices from Vehicles*. Washington, DC: U.S. Department of Transportation National Highway Traffic Safety Administration.

Greenberg, J., L. Tijerina, R. Curry, B. Artz, L. Cathey, D. Kochhar, K. Kozak, M. Blommer, and P. Grant. (2003). Evaluation of Driver Distraction Using an Event Detection Paradigm. Transportation Research Record: Journal of the Transportation Research Board, 1843, 1-9.

Groeger, J.A. (2000). Understanding Driving: Applying cognitive psychology to a complex everyday task. Taylor and Francis Inc., Philadelphia, PA.

Hancock, P. A. (2005). The tale of a two-faced tiger. *Ergonomics in Design*, 13(3), 23-29.

Groeger, J. A., and Clegg, B. A. (1998). Automaticity and driving: Time to change gear conceptually. In J. S. Rothengatter and E. Carbonell Vaya (Eds.), Traffic and transport psychology: Theory and application (pp. 137-146). Amsterdam: Elsevier.

Jenness, W. T., and Lattanzio, R. J., O'Toole, M., and Taylor, N. (2002). Voice-Activated Dialing or Eating a Cheeseburger: Which Is More Distracting during Simulated Driving? *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (592-596). Santa Monica, CA: Human Factors and Ergonomics Society.

Mattes, S. (2003). The Lane-Change-Task as a Tool for Driver Distraction Evaluation. In: Strasser, H.; Kluth, K.; Rausch, H.; Bubb, H. (Eds.): Quality of Work and Products in Enterprises of the Future (pp. 57-60). Stuttgart: Ergonomia Verlag.

Miyake, A. and Shah, P. (Eds.), 1999. Models of Working Memory: Mechanisms of Active Maintenance and Executive Control. Cambridge, UK: Cambridge University Press.

Monk, C., A., Boehm-David, D. A., and Trafton, J. G. (2004). Recovering from interruptions: Implications for driver distraction research. Human Factors, 46(4), 650-663.

Mortimer, R. (1993). The high-mounted brake light: A cause without a theory. *Proceedings of the Human Factors 37th Annual Meeting*, 955-959.

Mortimer, R. (1992). Weber's law and rear-end collisions. *The Michigan Academician*, 99-105.

Mortimer, R. (1990). Perceptual factors in rear-end crashes. *Proceedings of the Human Factors 34th Annual Meeting*, 591-594.

Nakayama, O, Futami, T., Nakamura, T., and Boer, E. (1999). Development of a steering entropy method for evaluating driver workload (SAE Paper 1999-01-0892). Warrendale, PA: Society of Automotive Engineers.

Netter, J., Kutner, M. H., Wachtscheim, and Wasserman, W., (1996). Applied linear statistical models. (Fourth Edition). New York: McGraw-Hill.

Norman, D. A., and Shallice, T. (1980). Attention to action: Willed and automatic control of behavior. CHIP Document No. 99. Centre for Human Information Processing, University of California, San Diego, La Jolla.

Noy, Y.I., and Lemione, T. (2000). Prospects for Using Occlusion Paradigm to set ITS Accessibility Limits. Presentation given at the 2000 Occlusion Workshop, Turin, Italy.

Posner, M.I., and Peterson, S.E. (1990). The attention system of the human brain. Annual Review of Neuroscience, 13, 25-42.

Recartes, M. A., and Nunes, L. M. (2003). Mental Workload While Driving: Effects on Visual Search, Discrimination, and Decision Making. .Journal of Experimental Psychology: Applied, Volume 9, Issue 2, June 2003, Pages 119-137.

Regan, D., and Gray, R. (2000). Visually guided collision avoidance and collision achievement. Trends in Cognitive Sciences, Volume 4, Issue 3, 1 March 2000, Pages 99-107.

Reyes, M., and Lee, J.D. (2004). The influence of IVIS distractions on tactical and control levels of driving performance. Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting (CD), 2369-2373. Santa Monica, CA: Human Factors and Ergonomics Society.

SAE (2004). Surface vehicle recommended practice: Navigation and route guidance function accessibility while driving. SAE J2364, August.

Shinar, D. (1998). Speed and crashes: A controversial topic and an elusive relationship. In National Research Council Committee on Guidance for Setting and Enforcing Speed Limits (Eds), Managing Speed: Review of current practice for setting and enforcing speed limits (Special Report 254) (pp. 221-276). Washington, DC: National Academy Press.

Shinar, D., Meir, M., and Ben-Shoam, I. (1998). How automatic is manual gear shifting? Human Factors, 40 (4), 647-654.

Stevens, A., Board, A., Allen, P., and Quimby, A. (1999, December). *A safety checklist for the assessment of in-vehicle information systems* (Report No. PA3536-A99). Crowthorne, Berkshire, England: Transport Research Laboratory (TRL).

Stuss, D. T., Shallice, T. Alexander, M. P., and Picton, T. W. (1995). A multi-disciplinary approach to anterior attentional functions. In J. Grafman, K. Holyoak, and F. Boller (Eds.), Structure and function of the human prefrontal cortex. Annals of New York Academy of Sciences, 279, 191-211.

Stutts. J. C., Reinfurt, D. W., Staplin, L., and Rodgman, E. A. (2001, May). The role of driver distraction in crashes. Washington, DC: AAA Foundation for Traffic Safety. (available in PDF at www.aaafoundation.org).

Tijerina, L., Kiger, S., Rockwell, T. H., and Wierwille, W. W. (October, 1996). NHTSA heavy vehicle driver workload assessment final report supplement -- Task 5: Heavy vehicle driver workload assessment protocol (DOT HS 808 467). Washington, DC: National Highway Traffic Safety Administration.

Verwey, W. B. (1991). Towards guidelines for in-car information management: Driver workload in specific driving situations. Report IZF 1991 C-13. Soesterberg, The Netherlands: Institutes of Perception.

Victor , T. (2004). Driving Support From Visual Behavior Recognition (VISREC) – Evaluations of Real-time Attention Support Functionality. Presentation given at the International Workshop on Progress and Future Directions of Adaptive Driver Assistance Research, Sponsored by the National Highway Traffic Safety Administration, May 13-14, 2004, at USDOT Headquarters.

Victor and Karlsson (in preparation)

Wickens, C. J. (1980). The structure of attentional resources. In R. S. Nickerson (Ed.), Attention and Performance VIII (pp. 23-257). Hillsdale, NJ: Erlbaum.

Wickens, C. D., and Hollands, J. G. (2000). Engineering psychology and human performance (Third edition). Upper Saddle, NJ: Prentice-Hall.

Young, K., Regan, M., and Hammer, M. (2003). Driver distraction: A review of the literature. Monash University Accident Research Centre (Report No. 206). Victoria, Australia: Monash University Accident Research Center.

Young, R. A., and Angell, L. S. (2003). The dimensions of driver performance during secondary manual tasks. Proceedings of “Driver Assessment 2003: The Second International Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design”, Park City, Utah. July 2003.

DOT HS 810 635
November 2006

U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

www.nhtsa.gov