88848

U.S. Department
of Transportation
**Federal Highway
Administration**

Publication No. FHWA-RD-94-089
December 1995

# Suggested Procedures and Acceptance Limits for Assessing the Safety and Ease of Use of Driver Information Systems

Research and Development
Turner-Fairbank Highway Research Center
6300 Georgetown Pike
McLean, Virginia 22101-2296

# FOREWORD

This report is concerned with developing measures and methods to assess the safety, easy to use and other driver responses to in-vehicle information systems. This report provides details on the measures and specifies test protocols. The purpose of this report is to:

(1) Identify measures of the safety of use, and driver comfort/convenience/confidence of in-vehicle information systems.

(2) Describe test protocols and dependent measures for assessing them.

(3) Identify levels of acceptance for the measures of interest.

This report will be useful to the designers of ITS-related driver information systems, scientists conducting automotive human factors research in academia, industry, or government agencies.

Sufficient copies of the report are being distributed to provide a minimum of two copies to each FHWA regional and division office and five copies to each State Highway agency. Direct distribution is being made to division offices.

Samuel C. Tignor, Ph.D., Acting Director
Office of Safety and Traffic
Operations Research and Development

# NOTICE

**Technical Report Documentation Page**

1. Report No.
FHWA-RD-94-089

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
SUGGESTED PROCEDURES AND ACCEPTANCE LIMITS FOR ASSESSING THE SAFETY AND EASE OF USE OF DRIVER INFORMATION SYSTEMS

5. Report Date
December 1995

6. Performing Organization Code
Account 080066

7. Author(s)
Paul Green

8. Performing Organization Report No.
UMTRI-93-13

9. Performing Organization Name and Address
The University of Michigan Transportation Research Institute (UMTRI)
2901 Baxter Road
Ann Arbor, Michigan 48109-2150

10. Work Unit No. (TRAIS)
3B3b-2032

11. Contract or Grant No.
DTFH61-89-C-00044

12. Sponsoring Agency Name and Address
Office of Safety & Traffic Operations R&D
Federal Highway Administration
6300 Georgetown Pike, McLean, Virginia 22101-2296

13. Type of Report and Period Covered
Final Report, 9/91 - 11/93

14. Sponsoring Agency Code

15. Supplementary Notes
This research was funded jointly by the Federal Highway Administration and the National Highway Traffic Safety Administration. The Contracting Officer's Technical Representative (COTR) was Nazemeh Sobhi (HSR-30).

16. Abstract
This report (1) identifies measures of the safety and ease of use of driver information systems, (2) describes test protocols for assessing safety and ease of use, and (3) identifies levels of acceptance. Only the driver interface is considered, not system safety considerations. Two protocols are described: an initial on-road test to assess the basic interface, and follow-on surveys at driver licensing offices after only small changes are made to the interface. The on-road test involves use of an instrumented car. From the data collected, measures of the standard deviation of lane position, mean speed, speed variance, the number and duration of eye fixations, and interface-specific performance measures (e.g., the number of turn error) can be obtained. For each measure, three levels of acceptance are specified: best expected, desired/planned, and worst case.

The measures listed above should be viewed as suggestions only. Normative data on driver performance are lacking, and the validity of the test protocols has yet to be established. There are also concerns about these procedures not being cost effective.

17. Key Words
IVHS, human factors engineering, ergonomics, navigation, driver information systems, test methods

18. Distribution Statement
No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
69

22. Price

Form DOT F 1700.7 (8-72)
(PF V2.1, 12/13/93)

Reproduction of completed page authorized

# PREFACE

## PROJECT OVERVIEW

The United States Department of Transportation (DOT), through its Intelligent Vehicle-Highway Systems (IVHS) program, is aiming to develop solutions to the most pressing problems of highway travel. The goal is to reduce congestion and improve traffic operations, reduce accidents, and reduce air pollution from vehicles by applying computer and communications technology to highway transportation. If these systems are to succeed in solving the Nation's transportation problems, they must be safe and easy to use, with features that enhance the experience of driving. A project was carried out to help develop IVHS-related driver information systems for cars of the future. This project concerns the driver interface, the controls and displays that the driver interacts with, as well as their presentation logic and sequencing.

The project had three objectives:

- Provide human factors guidelines for the design of in-vehicle information systems.
- Provide methods for testing the safety and ease of use of those systems.
- Develop a model that predicts driver performance in using those systems.

Although only passenger cars were considered in the study, the results apply to light trucks, minivans, and vans as well because the driver population and likely use are similar to cars. Another significant constraint was that only able-bodied drivers were considered. Disabled and impaired drivers are likely to be the focus of future DOT research.

A complete list of the driver interface project reports and other publications is included in the final overview report, 1 of 16 reports that document the project.[1] (See also reference 2 for an overview.) To put this report in context, the project began with a literature review and focus groups examining driver reactions to advanced instrumentation.[3,4,5] Subsequently, the extent to which various driver information systems might reduce accidents, improve traffic operations, and satisfy driver needs and wants, was analyzed.[6,7] That analysis led to the selection of two systems for detailed examination (traffic information and car phones) and contractual requirements stipulated three others (route guidance, road hazard warning, and vehicle monitoring) likely to appear in future vehicles.

Each of the five systems selected was examined separately in a sequence of experiments. In a typical sequence, patrons at a local driver licensing office were shown mockups of interfaces, and driver understanding of the interfaces and preferences for them was investigated. Interface alternatives were then compared in laboratory experiments involving response time, performance on driving simulators, and part-task simulations. The results for each system are described in a separate report. (See references 8 through 14.) To check the validity of those results, several on-road experiments were conducted in which performance and preference data for the various interface designs were obtained.[15, 16]

Concurrently, the contractor developed test methods and evaluation protocols, the contractor and a subcontractor developed design guidelines, and a subcontractor worked on the development of a model to predict driver performance while using in-vehicle information systems. (See references 17 through 20.)

Many of the reports from this project were originally dated May, 1993, the contractual end date of the project whereby reports were to be delivered. However, the reports were actually drafted when the research was conducted, more than 2 years earlier for the literature review and feature evaluation, and a year earlier for the laboratory research and methodological evaluations. While some effort was made to reflect knowledge gained as part of this project, the contract plan did not call for rewriting reports to reflect recent findings.

## THIS REPORT

This report is the second of two concerned with developing measures and methods to assess the safety, easy to use and other driver responses to in-vehicle information systems. In the first, the literature was reviewed, and measures of interest were identified.[17] This second report provides additional details on the measures and specifies test protocols.

This report is written for human factors practitioners given the task of assessing driver-related aspects of in-vehicle information systems. It should also be of interest to scientists conducting automotive human factors research. Those scientists may be working in academia, industry, or for government agencies. Within the Department of Transportation, both the Federal Highway Administration (FHWA) and the National Highway Traffic Safety Administration (NHTSA) are interested in this research, with NHTSA having expressed the greatest interest. Accordingly, this report emphasizes safety issues.

This report deals with a research area that is highly dynamic. Much of what is proposed here will require revision in the next few years. Specifically, this report describes two assessment protocols in detail along with levels of performance and suggests other test protocols. Much of what is proposed has received limited validation, and for many of the measures proposed, safe levels of performance are uncertain.

# SI* (MODERN METRIC) CONVERSION FACTORS

## APPROXIMATE CONVERSIONS TO SI UNITS

| Symbol | When You Know | Multiply By | To Find | Symbol |
|---|---|---|---|---|
| | | **LENGTH** | | |
| in | inches | 25.4 | millimeters | mm |
| ft | feet | 0.305 | meters | m |
| yd | yards | 0.914 | meters | m |
| mi | miles | 1.61 | kilometers | km |
| | | **AREA** | | |
| in$^2$ | square inches | 645.2 | square millimeters | mm$^2$ |
| ft$^2$ | square feet | 0.093 | square meters | m$^2$ |
| yd$^2$ | square yards | 0.836 | square meters | m$^2$ |
| ac | acres | 0.405 | hectares | ha |
| mi$^2$ | square miles | 2.59 | square kilometers | km$^2$ |
| | | **VOLUME** | | |
| fl oz | fluid ounces | 29.57 | milliliters | mL |
| gal | gallons | 3.785 | liters | L |
| ft$^3$ | cubic feet | 0.028 | cubic meters | m$^3$ |
| yd$^3$ | cubic yards | 0.765 | cubic meters | m$^3$ |
| NOTE: Volumes greater than 1000 l shall be shown in m$^3$. | | | | |
| | | **MASS** | | |
| oz | ounces | 28.35 | grams | g |
| lb | pounds | 0.454 | kilograms | kg |
| T | short tons (2000 lb) | 0.907 | megagrams (or "metric ton") | Mg (or "t") |
| | | **TEMPERATURE (exact)** | | |
| °F | Fahrenheit temperature | 5(F-32)/9 or (F-32)/1.8 | Celcius temperature | °C |
| | | **ILLUMINATION** | | |
| fc | foot-candles | 10.76 | lux | lx |
| fl | foot-Lamberts | 3.426 | candela/m$^2$ | cd/m$^2$ |
| | | **FORCE and PRESSURE or STRESS** | | |
| lbf | poundforce | 4.45 | newtons | N |
| lbf/in$^2$ | poundforce per square inch | 6.89 | kilopascals | kPa |

## APPROXIMATE CONVERSIONS FROM SI UNITS

| Symbol | When You Know | Multiply By | To Find | Symbol |
|---|---|---|---|---|
| | | **LENGTH** | | |
| mm | millimeters | 0.039 | inches | in |
| m | meters | 3.28 | feet | ft |
| m | meters | 1.09 | yards | yd |
| km | kilometers | 0.621 | miles | mi |
| | | **AREA** | | |
| mm$^2$ | square millimeters | 0.0016 | square inches | in$^2$ |
| m$^2$ | square meters | 10.764 | square feet | ft$^2$ |
| m$^2$ | square meters | 1.195 | square yards | yd$^2$ |
| ha | hectares | 2.47 | acres | ac |
| km$^2$ | square kilometers | 0.386 | square miles | mi$^2$ |
| | | **VOLUME** | | |
| mL | milliliters | 0.034 | fluid ounces | fl oz |
| L | liters | 0.264 | gallons | gal |
| m$^3$ | cubic meters | 35.71 | cubic feet | ft$^3$ |
| m$^3$ | cubic meters | 1.307 | cubic yards | yd$^3$ |
| | | **MASS** | | |
| g | grams | 0.035 | ounces | oz |
| kg | kilograms | 2.202 | pounds | lb |
| Mg (or "t") | megagrams (or "metric ton") | 1.103 | short tons (2000 lb) | T |
| | | **TEMPERATURE (exact)** | | |
| °C | Celcius temperature | 1.8C + 32 | Fahrenheit temperature | °F |
| | | **ILLUMINATION** | | |
| lx | lux | 0.0929 | foot-candles | fc |
| cd/m$^2$ | candela/m$^2$ | 0.2919 | foot-Lamberts | fl |
| | | **FORCE and PRESSURE or STRESS** | | |
| N | newtons | 0.225 | poundforce | lbf |
| kPa | kilopascals | 0.145 | poundforce per square inch | lbf/in$^2$ |

* SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.

(Revised September 1993)

## TABLE OF CONTENTS

## TABLE OF CONTENTS (continued)

# LIST OF FIGURES

# LIST OF TABLES

# INTRODUCTION

If a driver interface is to be safe and easy to use, it must be designed according to human factors principles, prototyped, and repeatedly tested. The last of the tests, the final assessment test, is the focus of this report, though preliminary tests are also addressed. Specifically, the purpose of this report is to:

1. Identify measures of the safety of, ease of use of, and driver comfort/convenience/confidence with in-vehicle information systems.

2. Describe test protocols and dependent measures for assessing them.

3. Identify levels of acceptance for the measures of interest.

As noted in the preface, given the Government agencies interested in this report (FHWA, NHTSA) and their mandates, comfort/convenience/confidence issues are given less attention than safety and ease of use. Further, less is known about comfort/convenience/confidence than safety and ease of use.

This report builds upon the review of measures and methods, the previous report in this series.[17] Readers unfamiliar with the human factors research on driver information systems should at least skim that document for context and constraints concerning the domain of this report. The application of this report is (1) Primarily for cars, essentially for five systems considered in this project (route guidance, traffic information, road hazard warning, car phones, vehicle monitoring), and (2) For only able-bodied, unimpaired drivers. Extension of this research beyond these limits is important, but it is beyond the scope of this research project.

It must be emphasized that it is very difficult to establish rigorous evaluation procedures. While it may be possible to abstractly identify test measures that are highly correlated with safety and ease-of-use in an absolute sense, schedules, funding, equipment, safety constraints, and personnel may rule out the collection of those that are ideal. Further, identifying what is unsafe or difficult to use assumes that one knows what is safe and easy to use. Defining what is safe or what is normal driving behavior is extremely difficult. As is discussed in this report, there is little normative data on driving behavior. Hence, establishing test procedures and acceptance levels at this time is likely to be heavily based on judgment and expert opinion without adequate validation. Similar views were expressed by Robertson and Southall in their survey of experts to determine if there was adequate information to establish a code of standard design practice for driver information systems.[21]

Measures, procedures, and acceptance levels detailed in this report should be viewed only as preliminary and incomplete. Those procedures and levels do not have sufficient support to be regarded as guidelines or standards. In that sense, the approach suggested is more of an exemplar than a specific approach to be followed. This report will be successful if it prompts further debate, research, and efforts to refine the proposed protocols. In fact, during various stages of drafting this report, terms such as strawman, and the genderless equivalent, strawperson, were used to describe the protocol proposed here.

To set the stage for establishing test procedures and acceptance levels, it is useful to know what constitutes a good test and how safety limits have been set in other contexts. Those issues are discussed in the following sections.

## WHAT IS A GOOD TEST?

The classic textbook on psychological testing identifies a good test as meeting at least six nonindependent criteria.[22] These criteria are similar but not identical to those used to select the individual performance measures for assessing driving information systems.[17]

1. Diagnostic or predictive value. The test identifies what is wrong or how well a person or product will perform a real task.

2. Standard procedure for administration and scoring. There is an agreed-upon way to collect the data and analyze the results.

3. Established norms. It is well known how well people in general (or in this case, systems) perform on the test.

4. Objective measure. The test can be scored with minimal human interpretation of the results.

5. Reliability. Repeated administration of the test to the same person or system under stable conditions yields the same results.

6. Validity. There are three types of validity: content validity (the test contains a representative sample of the behavior domain measured), criterion-related validity (the test predicts an individual's or system's behavior in specific situations), and construct validity (the test measures a theoretical construct or trait).

Many of the human factors tests described in the sections that follow do not satisfy all of these criteria. Until approximately 4 years ago, research on automotive human factors and driver interfaces had been at a low level for 20 years, so in many cases relevant human performance data are lacking. Where funding has been provided, it has been for crashworthiness, not crash avoidance. The lack of funding has had specific effects relating to each of the criteria named above.

First, lack of funding has constrained individual scientists to use equipment and protocols that they have on hand, rather than develop new methods. This makes it difficult for scientists to adopt procedures others have developed, even when they are superior, limiting opportunities for standard procedures to emerge. For example, a few organizations have driving simulators for sale. Wider use of them, with several organizations carrying out research on the same simulator, could lead to common test procedures. Without funds to purchase simulators, that standardization will not occur.

Second, normal driving behavior has not been studied. The focus has been on situations that lead to accidents. But to evaluate the impact of new systems and understand what is abnormal, comparison baseline data on normal driving is essential.

Third, reliability has not been adequately explored. In science, truth is established by multiple investigators independently examining some question. This is achieved through an open exchange of papers, reports, and dialogue at conferences, leading to a final conclusion. However, parallel funding of multiple organizations exploring the same question has historically been viewed as needless duplication. While not common in behavioral research, replication is essential as a basis for establishing the acceptability of a standardized test protocol.

Fourth, the exploration of validity has suffered as well. The pressure has been to find the answer and not to establish the validity of a variety of protocols (e.g., in simple laboratory tests, in driving simulators, on the road). For the purpose of specifying safety standards for in-vehicle displays, validity is a critical concern.

The IVHS community is pushing for answers to human factors questions now. It must realize that it will take many years to develop the test hardware and software, establish test procedures, and collect the necessary normative driving data.

Given these caveats, interim protocols for evaluating safety are proposed, though their validity needs to be established. In addition to these somewhat abstract concerns, the major influences in test selection are far more pragmatic—the cost of tests and the availability of equipment and software. Cost is critical to manufacturers and suppliers, who are concerned with burdensome regulations. One way to avoid such burdens would be to encourage the Government to allow the use of human factors tests, normally conducted during development, to substantiate that a system is safe and easy to use. This may reduce overall costs by minimizing required Government involvement in testing and certification. Unfortunately, these tests are nonstandard, with each manufacturer conducting evaluations in a manner particular to the resources and equipment on hand. Comparative data should also be provided for tasks involving non-IVHS interfaces that are generally accepted as presenting acceptable risk to the driver (e.g., operating the lights or wiper). The interpretation of the data could be highly judgmental and the basis for such judgments is uncertain. Hence, a standard formal assessment protocol is desired.

Further, decisions concerning the selection of an evaluation begin with the selection of an experimental context that provides a valid, replicable, accessible, and cost-effective assessment of the issues of interest. It does not begin with an abstract consideration of potentially appropriate dependent measures. Decision criteria for the selection of evaluation contexts and the dependent measures appropriate for each context follow.

This report describes, in detail, on-road tests, to which each new system should be subjected, and driver-licensing-office tests, which are appropriate for existing systems that are being modified. Other types of tests may be acceptable and are discussed later. However, the author believes on-road testing should be the primary determinant of whether or not a system should be considered safe and this opinion is shared by others.[23] The rationale for this view is described in great detail in other sections of this report, and to a large degree is based upon research conducted as a part of this project. In brief, while a great deal can be simulated, there has been minimal research to validate how driver behavior in simulators matches driver behavior in the real world. (Some validation research on this will also be conducted by the contractor over the next year and is reportedly being conducted at the University of Iowa.) The lack of validation studies is most problematic for navigation system evaluations where traffic density, landmarks, and highway signs are all critical. At the present time, representing all of these items (dense traffic, landmark details, and signs readable at a great distance) is very challenging. For systems that involve considerable risk during testing, such as collision avoidance, the use of driving simulators is preferred.

For navigation systems, in-vehicle information is critically linked to information presented on highway signs. At the present time, except for the FHWA simulator, there are no simulators for which simulated signs can be read at visual angles similar to those on the road. It is likely to be some time before computer-generated image systems with the desired capabilities will be available in the automotive industry.

Another problem is cost. Instrumented cars rent for about $100 to $200 per hour plus fuel and personnel costs, while high-fidelity simulator costs are typically $1000 to $2000 per hour. (In estimating the cost of a car, simple comparisons may only include the vehicle cost and not the instrumentation cost.) For the 80 h of subject time that might be needed for a test, an instrumented car would cost $8,000 to $16,000, while the simulator would cost $80,000 to $160,000. An instrumented car costs $140,000 or more to develop, depending on the equipment installed. There are also significant simulator-induced motion sickness problems to overcome, problems that are common to older drivers, the group of greatest interest. Further, additional time is required for installation of the driver interface to be examined. In the future, simulators may become a more attractive option as their costs decrease and capabilities increase.

It is not necessary or cost effective to conduct a full-scale on-road test each time an interface is modified to assess safety or ease of use. If the change is "minor," for example revisions of a few abbreviations for switch labels, retesting of the entire interface is likely to be wasteful. For this reason, manufacturers in those cases are provided with the option of conducting simple "paper and pencil" surveys of interface nomenclature and graphics preferences/interpretations at a local driver-licensing office. Other part-task simulations could be substituted as well.

Another dilemma concerns the tests of ease of use. Traditionally, the Government has emphasized the safety of motor vehicles, not their usability. But if IVHS is to succeed, products must be easy to use. Accordingly, usability acceptance levels are offered.

Where possible, multiple levels of acceptance are provided for each test. In reality, the final product design is the result of a series of trade-offs. Having only a single, minimum level of acceptance as the only acceptance level leads to systems uniformly minimal in their acceptability. Because it only has minimum acceptance levels, the best known human engineering standard, Military Standard 1472, has been criticized.[24,25] To provide flexibility in design, some tradeoffs should be allowed. If an interface does poorly (but is minimally-acceptable), for one criterion considered independently of others, it should be required to do well, or above minimally, on others.

## WHICH ORGANIZATIONS SHOULD CONDUCT THE TESTING?

There are three choices as to who should conduct assessment tests:

1. Independent test laboratories.
2. The Government aided by contractors working directly for them.
3. The manufacturers augmented by contractors; the work is subject to Government review.

Considering limitations in the availability of personnel and equipment to conduct tests, the criteria given below should be used in selecting among testing organizations. These criteria are different than those used for selecting measures because the issues are different.

Minimize:

- The opportunities for falsifying test results.
- The cost of tests.
- The time required to conduct tests (interference with production schedules).

Maximize:

- The ability to use the data to improve product design.
- Accessibility of the test facilities to industry and Government.

The author's interpretation of how test options fare with regard to these considerations is shown in the table that follows. No formal weighting scheme was applied in evaluating these criteria, and the process of selection of the test organization was not formalized.

Table 1. Decision criteria.

| Criteria | Independent labs | Government | Industry |
|---|---|---|---|
| People and equipment | Lacking (see text) | Lacking (see text), could lead to more bureaucracy | OK in some cases |
| Falsification | Not an issue | Not an issue | Not a problem if checks provided |
| Cost | Unknown | Unknown | Unknown |
| Duration | Moderate | Long | Short |
| Results lead to improved designs | Unlikely | Unlikely | Should occur (could piggyback proprietary tests to protocol) |
| Access to facilities | Not good | Not good | Good |
| Product confidentiality | Fair | Uncertain | Good |

Readers should note that considerations of validity, reliability, and accuracy of the results are not listed as criteria. It is expected that all organizations will provide high-quality data, and there is no prior reason to expect differences between organizations.

The independent test laboratories (e.g., Underwriters Laboratories) have a long tradition of conducting safety tests. Their independence tends to isolate them from the manufacturers, making it difficult to feed back the results from tests that might lead to improved designs.

The independent laboratories are generally held in high regard. To maintain that image, there is strong pressure not to falsify results. Many people will not buy small electrical components such as motors, switches, and relays or, consumer products such as blenders and toasters that lack a UL label. These organizations are used by both large and small manufacturers, but in particular the smaller ones that do not have extensive in-house test facilities.

Most of the independent laboratories do not have the facilities to test large products such as cars, and those that do, such as the Consumers Union, may find that routine assessment testing will detract from their main mission of expanding consumer awareness. For the necessary evaluations of advanced driver technologies, such laboratories at the present time do not have cars with the desired instrumentation, nor has there been any discussion of obtaining driving simulators that might be required for future work. For these reasons, it is unlikely independent test laboratories will be involved in driver interface assessment and are not considered further.

Even more so than for the independent laboratories, there is no motivation for the government to falsify assessment test results. Government laboratories have traditionally been very independent of the manufacturers, in part (and especially within DOT), because research is conducted to support regulatory activities of the agency. This inhibits feedback of results to manufacturers, as it becomes difficult to separate suggestions for design improvements from required changes.

Presently, the Government does not have sufficient resources to conduct assessment tests. Safety regulations regarding in-vehicle displays fall under the purview of Federal Motor Vehicle Safety Standard 101, Controls and Displays, a statute that is the responsibility of the National Highway Traffic Safety Administration (NHTSA). However, NHTSA's primary role has been as a contract monitor for research related to this subject. While its scientific personnel know how to conduct such tests, NHTSA lacks the human factors support personnel (to test subjects, to maintain human factors laboratory equipment and test vehicles) and equipment (vehicles instrumented for human factors evaluations, driving simulators, etc.) needed. The Federal Highway Administration's (FHWA) Turner-Fairbank Highway Research Center has been engaged in on-road studies of driver behavior and is in a better position to conduct such tests. It is in the process of acquiring an instrumented car suitable for the required testing. The driving simulator at Turner-Fairbank, while useful for research, may not have the fidelity needed for assessment tests, but the proposed National Advanced Driving Simulator (NADS) in Iowa could be suitable.

Given current staffing limits, the Government could augment its personnel with contractors who would conduct the tests. However, the lack of Government test facilities may necessitate off-site evaluations, reducing oversight of the assessment test.

Resources to conduct assessment testing exist in industry, and their capabilities are expanding. Information in the section that follows on the status of industry capabilities comes from the author's conversations with a wide variety of individuals in industry who wish to remain anonymous. The author is in constant contact with automotive human factors work, both professionally and personally, and believes his impressions are reasonably well informed. Public information on capabilities does not exist.

To make decisions about the selection of assessment techniques, the availability of facilities in likely vendors needs to be considered. As an example, at the time this protocol was developed, General Motors Corporation (GM) was in the strongest position with a moderate number of human factors specialists (mostly in the Systems Engineering Division, some at Delco), several instrumented vehicles, and two high-fidelity simulators in Warren, Michigan and a third at the Hughes facility in El Segundo, California. GM had a usability laboratory in Warren and other facilities. The Ford Motor Corporation had a human factors group of approximately 15 people on the Design Staff with others in the Electronics Division. They had space available for laboratory experiments (but not a dedicated laboratory) and the Scientific Laboratories has just completed a driving simulator. Chrysler Corporation had three human factors specialists, no instrumented cars, and was developing a driving simulator. Because of its limited staff and inexperience with on-road human factors tests, Chrysler would have probably contracted out the assessment testing. Suppliers would also have contracted out some of the testing (as Motorola Corporation has done in the ADVANCE project). Others, such at American Telephone and Telegraph (AT&T), with large human factors groups, would have conducted the testing themselves. Those organizations without adequate facilities and personnel would either have a university that has had an ongoing relationship with DOT (e.g., University of Michigan through the Displays project, FAST TRAC, and intelligent cruise control efforts, Virginia Polytechnic Institute through the fatigue project, University of Iowa through the National Advanced Driving Simulator) conduct the testing, or a private company (e.g., Systems Technology, Inc., involved with intelligent cruise control).

For European manufacturers, access to instrumented vehicles, simulators, other test equipment, and the human factors people to operate them is not a problem. This is the result of interchanges between organizations involved with the DRIVE and PROMETHEUS programs. Daimler-Benz Gmbh (Mercedes) and Bayerische Motoren Werke (BMW) had both simulators and instrumented vehicles, along with support personnel. Fiat Group had a few human factors people and had just completed a simulator. Less is known about Volkswagen (VW), French, or British manufacturers. Where contract support is needed, these companies often turn to Swedish Traffic Institute (VTI), Institute for Perception (TNO), Institut National de Recherche sur les Transports el Leur Securite (INRETS-the French transportation safety research institute), the University of Groningen, and others for assessment testing. These organizations had the vehicles, simulators, and personnel required for safety and usability assessments.

For Japanese manufacturers, the equipment and facilities necessary for safety and usability assessments were readily available, but the Japanese manufacturers had less experience in conducting formal human factors tests than their American and European counterparts. That difference will disappear in the near future. Virtually all of the manufacturers had instrumented cars and many had driving simulators (Mazda Motor Corporation, Nissan Motor Company, Toyota Motor Corporation, Mitsubishi Motors Corporation). Little is known about Japanese suppliers (Sumitomo Electric Corporation, Matsushita Electric Corporation), though based on tradition, they would have conducted joint tests with the manufacturers.

Falsification of results, an argument against manufacturer testing, is unlikely to be a problem. Currently manufacturers perform crash and emissions tests of their own products, though the Government reviews reports and audits tests to check manufacturers' results. Additional checks could come from public distribution of test reports, allowing manufacturers and suppliers to check each other. Normal competitive pressures are likely to cause careful scrutiny of test results, and when procedures are improper, protests are likely. The manufacturers are not under much pressure to falsify results because many of the IVHS products will initially be options, though profitable ones.

The extent to which test results can lead to improved products is an important consequence of industry testing. It is quite likely that design personnel will witness assessments and will receive preliminary results well before the technical report describing the results in detail is available. This promotes rapid improvements in interface design.

The final item to consider is keeping proprietary product details confidential, especially prior to production. Knowing that a competitor is going to offer, for example, a navigation product, with particular features, is a significant competitive advantage. Manufacturers are very reluctant to show new products to anyone not directly involved with them, either inside or outside the company. This consideration favors companies testing their own products.

Hence, (1) the availability of people, equipment, and facilities, (2) checks and balances against falsification of results, and (3) the potential that assessments could result in improved designs leads to recommending industry-based assessment testing for safety and ease of use. In addition, because testing is in-house, scheduling flexibility is likely to be greater, an important consideration for fast paced product development schedules.

Readers may note that cost issues have not been addressed. Cost figures are not available for any of the target organization categories. Therefore, cost cannot be applied as a differentiating selection criteria here.

## WHAT QUALIFICATIONS MUST THE STAFF HAVE?

To carry out assessment tests, an interdisciplinary team will be required, though all of the knowledge useful in conducting these tests may not be present in members of the team. When staffing gaps are present, qualified individuals should be co-located with the evaluation team and not separated by organizational or administrative barriers (e.g., requiring permission of a supervisor before providing advice). The number of personnel and the roles they play will vary from organization to organization. Among the personnel involved with tests will be engineers, technicians, programmers, user-interface designers, and statisticians. Critical expertise includes:

- General knowledge of automotive engineering.
- Experience in testing human subjects.
- Experience in the statistical analysis of large data sets relating to driving.
- Knowledge of human factors as it relates to driving and familiarity with the measures being collected.
- Instrumentation specialists.
- Knowledge of Government standards and requirements.

There is a clear need for a human factors professional to play the lead role in conducting the assessment test. These individuals may be human factors scientists, human factors engineers, or experimental psychologists experienced in automotive human factors. A major difficulty is to determine if the team leader is qualified. There is no single, generally accepted certification process and there are very few alternatives. One way would be to require that this individual be a Member of the Human Factors and Ergonomics Society (HFES), the Ergonomics Society, be certified by the Board of Human Factors Professionals (BHFP), or have equivalent training. Additional information on HFES and BHFP can be obtained from the Human Factors and Ergonomics Society (Box 1369, Santa Monica, CA, phone: 310-394-1811). Information concerning the Ergonomics Society can be obtained by writing or calling them (Devonshire House, Devonshire Square, Loughborough LE11 3DW, United Kingdom, phone: 44 509 234904). Membership in these organizations does not guarantee an individual will be well qualified, only that they will have at least minimum knowledge of the subject matter.

To be a Member of the Human Factors and Ergonomics Society (versus an Associate or Affiliate Member) requires a bachelor's degree and a minimum of 5 years of full-time experience in the profession. Up to 4 years of academic experience may be substituted for work experience. Applicants must have their application forms endorsed by two members of the Human Factors and Ergonomics Society, who are also responsible for establishing the experience of applicants. Applications are reviewed and checked by the Society's membership committee.

It is also important that individuals have a general knowledge of automotive human factors. Given the shift from defense to civilian applications, it is likely that some candidates for lead role are human factors engineers working for defense contractors. It will take some time for them to be retrained in automotive engineering.

While it is important that participants in assessments be well qualified, the author does not believe at this time it is necessary to develop detailed selection criteria for identifying qualified individuals, panels to review the criteria, and procedures for weighting responses. Given the resources available, it was felt most appropriate to focus on the test protocol. If necessary, personnel selection criteria can be added at a later date.

An important qualification is expertise in testing human subjects. Often those who test subjects in industry have a bachelors degree in psychology, but they could be engineers. They are often junior personnel, and in the case of a university, are typically students. It is essential that the test leader have expertise in this area as well, which is presumed for Members of HFES or those that are BHFP-certified.

There is a trend to try to bring designers closer to the users of products. One ramification of that trend is to have designers carry out the usability tests and test subjects. When designers observe usability problems directly, they are more likely to make improvements to the design than if they only receive a report describing problems. While that may be appropriate in the early phases of design, for assessment tests, use of designers as subject runners should only occur when designers have had prior experience in subject testing, and when they are supervised by a human factors professional.

Finally, it is important that the team have statistical expertise, which in many cases might be a second role for the human factors professional. In some organizations this expertise is in an in-house statistical consulting group. The data from on-road tests can be voluminous, and because it is real-world data, some of it is likely to be missing or in error. Data reduction and filtering is very important. The number of people with experience in analyzing driver-performance data is very small.

It must be emphasized that the pool of talent to conduct assessment tests is limited but recent increases in funding for research on driver interfaces should expand the pool.

## HOW SHOULD THE TEST RESULTS BE PRESENTED?

Sufficient detail should be provided such that the work is publicly replicable; that is, the method and results can be duplicated within the limits of statistical error. The assessment test should be contained in a detailed technical report containing an abstract, a brief introduction identifying the issues of interest, a test plan (description of the subjects tested, test equipment and materials, test activities and their sequence), results, conclusions, and perhaps references and an appendix. Among the items that must be included are screen dumps or clear pictures of the various screens, the text for auditory messages, and a detailed map showing the test route. Based on the author's experience, it should be possible to thoroughly describe the test and the results in less than 60 pages. If the test protocol could be standardized, then it could either become an appendix to the report or referenced, so the body of the test report could be only 25 to 40 pages in length. The report describing the final on-the-road experiment of this project was 99 pages long, 50 pages of which were results.[16] Standardizing the protocol and examining a more constrained topic should reduce the number of pages required to describe the procedure and results. The greater the number of pages expected, the more expensive the test becomes. To provide a sense of the detail desired in the report, readers are encouraged to look at articles in Human Factors, the journal of the Human Factors and Ergonomics Society. It must be emphasized that a journal article is not the desired end product of assessment tests. It would be desirable if report generation could be a form-filling exercise.

User interfaces are highly dynamic. It is difficult to get a sense of how an interface works on paper. To provide the desired documentation, a videotape demonstrating the operation of the interface should be among the materials submitted. If it is submitted, then only a summary of the transitions from one screen to another is needed in the written report. It is likely that a rigidly mounted camcorder will provide an image of acceptable quality. The desired footage may also be obtained from cameras in instrumented cars. A broadcast-quality tape exceeds what is needed. The author's experience has been it costs less and take less time to document screen sequencing and user interactions using a videotape than in a written report.

The use of videotapes to supplant text and figures in technical reports is clearly a novel approach. Certainly duplicating a videotape is much more time consuming and costly than photocopying printed text. However, this difference, even for several hundred copies, is much less than the cost difference between producing a short report and producing a short videotape ..It has been the author's experience that 10-minute videotape produced with "home" quality equipment is generally more informative, in terms of understanding and appreciating the functioning of an interface, than 20 pages in a carefully written technical report.

In the near future, submitting video clips on disk as part of a multimedia report may also be an option. If the PowerPC platform is successful, limitations associated with operating system-specific multimedia software may disappear.

Assessment tests may involve the collection of both proprietary and nonproprietary data. If the proprietary aspects of the research involved examining other interfaces designs, the report submitted could describe them only very generally ("another voice-based interface). If the proprietary aspects concerned additional measures not required in the test protocol, report authors could refer to "other proprietary measures" in the protocol description and not identify findings related to them in the results section of reports submitted to the Government. As an example, in the TravTek project, both proprietary and nonproprietary data were collected by GM and there have been no reports of problems of keeping the data separate.

## HOW HAVE SAFETY LIMITS BEEN SET IN OTHER CONTEXTS?

Safety limits have been set for many products, services, systems, and tasks. To put setting such limits for in-vehicle displays in context, it is useful to consider how this is done for other applications. There are many examples. The Federal Aviation Administration, a unit of the Department of Transportation, certifies aircraft and cockpits as being air worthy, which is a safety assessment. Drugs are certified as being safe by experiment. Medical devices are beginning to be certified as being safe to use in the home. Nuclear power plants are certified by the Nuclear Regulatory Commission. There should be significant benefits in formally examining the human factors assessment procedures used by these organizations.

One interesting example with which the author is familiar is the development of standards for safe two-handed lifting in the National Institute for Occupational Safety and Health (NIOSH) Work Practices Guide.[26] While this application is quite removed from automotive human factors, it represents a context in which ergonomics guidelines have been well accepted. The application of the NIOSH guidelines has had a significant impact on human health and safety by reducing overexertion injuries, especially back injuries related to lifting. Particularly noteworthy of these guidelines is the solid scientific basis for them.

Four classes of criteria were used to set NIOSH lifting limits: epidemiological (health and injury statistics), biomechanical (musculoskeletal analysis), physiological (metabolic and circulatory responses), and psychophysical. In the biomechanical analysis, the bones and muscles were treated as mechanical links and sources of force, with objects being lifted in the hands (at some distance) leading to forces and moments operating on the L5/S1 disc (fifth lumbar/first sacral) in the lower back. The primary question is whether the imposed load is sufficient to cause a spinal disc compression failure. Such levels can be readily measured (for example, by taking spinal sections from cadavers and subjecting them to a mechanical compression test).

For lifting which is continuous, the limits are set by aerobic capacity. The NIOSH Guide describes several studies of aerobic output for adults. Based on that research, aerobic limits were selected to avoid heart rate creep (a heart rate that continually climbs over the day). Also described are equations that relate the physical characteristics of lifting (weight of the object, height of the lift origin and destination, etc.) to the estimated metabolic energy expenditure for that task.[27]

Finally, the NIOSH Guide also presents data for occasional high frequency lifting where the constraint is local muscle fatigue. For this category, the limits are based on percentile data for dynamic and isometric strength from a variety of studies.

In recognition of the limits of the data and human variability, the NIOSH Guide contains two sets of limits, the Maximum Permissible Limit (MPL) and the Action Limit (AL). For each of these two limits there are specific maximum values L5/S1 compression forces (in pounds), metabolic rates (in Kcal/min), and weights to be lifted (in pounds).

What has made the NIOSH Lifting Guide a success?

1. The task characteristics are well known and few in number.
2. The task characteristics are readily quantified (weight of the object, frequency of lift, etc.).
3. The mediating indicators (compressive forces, aerobic capacity, muscle forces) are well known, and few in number.
4. The mediating indicators are readily quantified using well understood measurements. Muscle and disk forces can be estimated from a biomechanical analysis of the weight of the object lifted, its relative location, and other task characteristics. For metabolic demands, the energy required to move weights various distances emerges from physics (energy required to move mass) and biochemistry (the energy liberated when oxygen is combined with various foodstuffs to yield carbon dioxide, water, and other byproducts).
5. There are a limited number of well-defined ultimate performance measures (the number of injuries and their mean severity).
6. The relationship of task characteristics, mediating variables, and injury mechanisms is well known.
7. In setting criteria, it was recognized that a single threshold would be unacceptable, hence multiple limits representing varying degrees of severity were established.

These criteria are not currently met for driving. While the task characteristics are large in number (variables associated with traffic, variables associated with the road, vehicle dynamics, etc.) many can be well quantified. However, lifting tasks are deterministic and repetitive (with short cycles) while driving is probabilistic. In the case of driving, the list of mediating variables (speed variance, lane variance, time-to-line crossing, etc.) is quite large. While some variables can be determined from physical laws, many of them cannot. While the linkage of them with task characteristics can be achieved, it has not been in many cases for normal driving. As an example, for daytime driving on a straight road at 89 km/h (55 mi/h) in a 3.7-m (12-ft) lane with average headways of 30.5 m (100 ft), the lane or speed variance for a 45-year-old driver cannot be predicted. Fortunately, the ultimate performance measures are well known and few in number (number of crashes per mile driven, fatalities per mile driven, etc.). However, the connection between these measures and the driving performance measures is not well quantified. Accordingly, it will be very difficult to establish rigorous safety standard for in-vehicle displays.

## HOW SHOULD SPECIFIC SYSTEMS BE TESTED?

The United States approach to governmental safety assessment of products is to test the final version. Based on some measure of effectiveness that is linked to accidents, the product is deemed either safe or unsafe. This is in contrast to the European approach for some matters related to occupational safety (especially in the United Kingdom), where the process for developing the product or service may be reviewed by government agencies, in addition to the certification of the final outcome (the product or service).

As was noted earlier, the success of IVHS products depends not only on those products being safe to use, but they should be easy to use and enhance the enjoyment of driving. Human factors experience has shown that a final check of a product (either for safety or ease of use) will have only a small impact on how safe and easy to use a product is; further, only minimum standards are achieved. It is therefore appropriate to consider mechanisms that incorporate safety and usability into the design process.

Philosophically, this view is embodied in the Gould and Lewis Design Principles, which the computer industry has applied with great success to user interfaces.[28] Depending upon which paper by Gould and Lewis one reads, either three or four principles are cited. The three main principles are (1) early focus on users and tasks, (2) empirical measurement, and (3) iterative design. Certainly, the points Gould and Lewis make and the contractual objectives (and tasks) of this project have a great deal in common. "Early focus on users and tasks involves direct study of their cognitive, behavioral, anthropometric, and attitudinal characteristics, and in part by studying the nature of the work expected to be accomplished."[25] The key to successful application of this principle is direct contact with users carrying out real or representative tasks. Reading studies is not sufficient. In this case, both the literature reviews completed at the beginning of this project and the focus groups were driven by that principle.[3]

The second principle, empirical measurement, directly concerns the focus of this report—attempting to measure what people do. There are two aspects to this principle: actual behavioral measurements of learnability and usability, and conduct of these tests very early in the development process. Early design tests may involve very crude prototypes evaluated in a laboratory. Because of the simplistic nature of such tests, correlations with safety of the performance measures may not be high. However, such tests have a significant impact on the safety and ease-of-use of a product. It is only in early design stages that testing can have a major impact on the final product design.

Related to this point are the types of measurements to be collected. Design measurements, which specify how something should be made, are most appropriate for mature products or where physical interfacing is required. For example, bumpers will be most effective if the heights of the colliding bumpers are the same. In performance-based requirements, how well something is done (e.g., the time to read a display) is specified. User performance specifications provide manufacturers with more flexibility than design measurements and are therefore generally preferred as a means for specifying driver interfaces.

Iterative design is the third principle. Safe and easy to use products are the result of repeated cycles of design, testing, redesign, and retesting. The quality of the product is proportional to the number of times through the design-test loop, though there are diminishing returns. Iterative design should end when performance requirements are met.

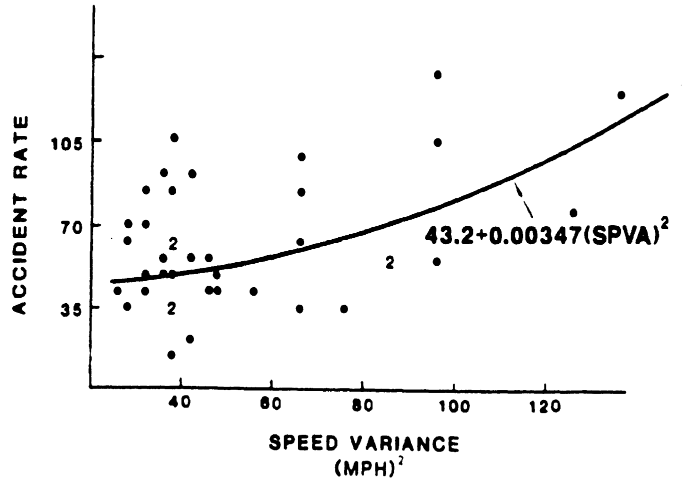
A second key philosophical point that has emerged from the human-computer interaction literature concerns the setting of pass-fail criteria. At Digital Equipment Corporation (DEC), there are three levels of pass-fail criteria, instead of one: best case, planned level, and worst case.[29] In all instances the current level of performance of the current product for each attribute or measurement of effectiveness is known. Performance levels for competitive products are also usually known. Typically, the worst case is the same as the present level for the current product, the planned level is as good as the best of the competition (or some reasonably achieved level), and the best case is better than the best of the competition by a significant amount. In the automobile industry, such targets are set in so-called best-in-class or BIC evaluations and are a key part of quality-function-deployment (QFD) analyses. (See references 30 through 34.)

DEC engineers and scientists have found that real design represents a series of tradeoffs and that if all of their standards are minimums, then there is no incentive for products to be designed to achieve higher levels of performance. However, the key to long-term success is to market the best possible product at the lowest cost. In addition, products should be continually improved, key principle in contemporary Japanese engineering methods and one reason their products have been successful.

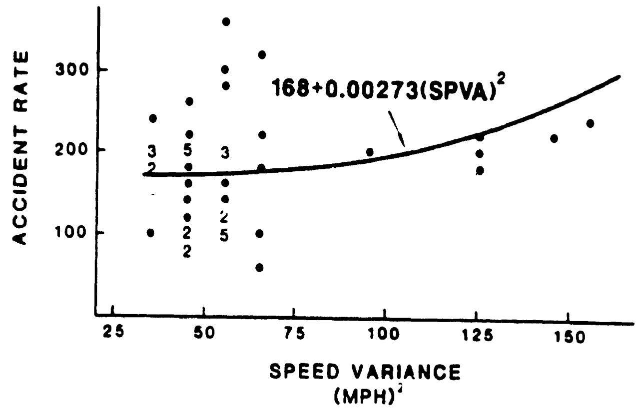
As was mentioned earlier, common human engineering practice is to meet the requirements of Military Standard 1472.[22] Because the standard gives only one value, a minimum (for knob spacing, access holes, etc.), most tanks, planes, and ships often just barely meet the minimum standard. When minimums are set for a particular parameter, it is generally assumed that other parameters are at typical or normal values. So, for example, in setting a requirement for the minimum contrast of a character, it is generally assumed that the character luminance is at a normal value, not a minimum value. The conflicting minima problem commonly occurs for anthropometric criteria. If a task requires 5th percentile grip strength and a 95th percentile reach, then more than 5 percent of the sample will not be able to complete the task. Since the two measures are not perfectly correlated, some people will not have enough strength, others not enough reach, and some will be lacking on both dimensions. Hence, not everyone will be accommodated.

Finally, in identifying levels of safety, there is some arbitrariness in identifying levels as representing safe and unsafe conditions. For example, figure 1 shows experimental data concerning the relationship between speed variance and accident rate. Clearly, adding driver-information systems could distract the driver and could cause him or her to pay less attention to maintaining a constant speed. However, there is no clear point at which a change in the speed variance causes a dramatic increase in accident rate. It is for these reasons that binary safe or unsafe criteria are not provided in this report. Instead, varying degrees of safety are identified.

It is not easy to determine how a procedure might be implemented to certify the process for developing a safe and easy-to-use driver interface. With regard to Gould and Lewis principle 1, (early focus on users and tasks), a manufacturer might be required to submit a report describing the users and what tasks are to be carried out using the product. With regard to principles 2 (empirical measurement) and 3 (iterative design), the manufacturer could be asked to submit reports describing the tests carried out during development to determine product safety and the acceptance levels of various measures. However, in discussing such ideas with the sponsor, the research team was left with the impression that the Government was not interested in an approach that left specification of the test protocol up to the manufacturer, and relied upon ad hoc tests conducted during product development.



a) Interstate highways.



b) Arterials

Figure 1. Speed variance versus accident rate.[35]

The advantage of this unstructured approach is that it utilizes information that manufacturers are (or should be) collecting now and does not burden them with additional tests. Second, it emphasizes the development process, which will ensure that the final product is safe and easy to use. In the future, some of that material may be needed to satisfy the requirements of ISO Standard 9000 (quality).

However, at this time, the disadvantages of reviewing the design process may outweigh those advantages. Potentially, manufacturers would be required to release proprietary marketing data in the report on users and their tasks. The manufacturers would not want to disseminate such data even after a product is on the market. Since the iterative design tests carried out vary from manufacturer to manufacturer and the approaches followed in those tests are loosely connected with on-the-road usage, establishing the adequacy and appropriateness of a test would be very difficult. Because there are so many potential approaches, it would be difficult to assemble enough data to establish the validity of each approach. A cost-effective alternative would be to have a panel of experts review each approach using a predefined set of criteria. However, there are no empirical data to indicate how well expert review of design processes correlates with performance of the final product.

At best, the Government can review the final product, not how the product was designed. Accordingly, a conventional test protocol for assessment testing is given in the sections that follow. However, anything the Government can do to encourage manufacturers to follow the Gould and Lewis principles can only be beneficial. The products designed should be safer and easier to use—the desired outcome.

# ON-THE-ROAD TESTS

## PURPOSE/ISSUE AND SYSTEMS TARGETED

The purpose of this procedure is to test a wide variety of systems. It is intended as the primary safety and ease of use assessment procedure. Interfaces tested may be working models or prototypes if they differ in inconsequential ways from real products.

For the moment, there is no requirement to examine activities not performed while driving, that is, predrive activities such as destination entry. The ultimate success of some systems, such as navigation, will depend upon successful completion of predrive activities. However, for this first attempt at developing an assessment protocol, activities that do not have an impact on safety or usability while driving have been excluded. Clearly, an exhaustive evaluation of a driver interfaces (the controls and display with which a driver interacts along with the associated system logic) should include pre-drive activities. Manuals and audio-visual instructional materials should also be reviewed.

In tests carried out on public roads, it is important to assure that prior laboratory tests have been carried out to ensure that prototypes have some reasonable measure of safety and that participants in experiments are not exposed to unnecessary risk. Such laboratory tests are not specified here.

## HUMAN SUBJECTS REVIEW

On-the-road tests must satisfy the constraints of the local Human Subjects Review Panel. Organizations that lack Human Subjects Review Panels should either establish one (following guidelines of the Department of Health and Human Services) or consider utilizing the panel of a local university. If the university has a medical school or psychology department that does human research, it should have such a panel. Panels are interdisciplinary, with members often being drawn from psychology and medicine (for their technical backgrounds), theology (for an ethical perspective), and other departments. Members of the review panel may be drawn from outside the organization conducting the research.

Constraints on testing human subjects are well known to people experienced in conducting human factors tests, the target audience for this report. For those without a human factors or psychology background, these panels are provided with forms giving a brief synopsis of the research and how subjects are to be recruited. For a typical human factors experiment in the author's organization, this description is approximately one page long. A subject consent form is also included. Concerns of the review panel in terms of the treatment of subjects are deception, punishment, use of drugs, covert observation, interviewing of children, mental or physical stress, procedures which risk physical harm, socially unacceptable materials, and the invasion of privacy. The panel has the options of approving the research plan as submitted, asking for modifications, or declining to provide approval.

## EQUIPMENT, MATERIALS, SOFTWARE

Field tests involving IVHS technologies of the type proposed here require instrumented cars. The product—or a simulation of it—should be installed in either a preproduction or current vehicle, or one whose driver interface is substantially unrevised since the last model changeover. (Model changeovers currently occur every 4 to 5 years.) The instrumentation required depends upon the detail of the evaluation. For the purpose of the on-the-road evaluation proposed here, the vehicle should contain a camera and VCR to record the forward scene, equipment for recording the location and duration of each eye fixation, equipment to record forward speed [at 10 Hz to the nearest 0.16 km/h (0.1 mi/h)], and equipment to record lateral position [to the nearest 33 cm (0.1 ft) at 10 Hz]. This is similar to the equipment suite in the TravTek camera car and the UMTRI instrumented car.[36] The PATH project in California may also have vehicles that contain representative instrumentation.

The system to be tested can be either a real product or a simulation. The author's experience with on-the-road use of SuperCard simulations of route guidance and other systems suggests that, when properly implemented, drivers do not realize the test interface is simulated. Simulations in HyperCard, Toolbook, Plus, and other applications should perform equally as well as real products. SuperCard, HyperCard, and Plus are Macintosh applications, though Plus code also runs on IBM-compatible computers.[37] Toolbook is an IBM-compatible application that runs under Windows. There may also be suitable applications that run on Amiga computers, though in the United States, use of Amiga hardware for interface prototyping (or engineering applications in general) is rare. While there are other platforms that support interface prototyping (e.g., Silicon Graphics), the hardware is currently too large to fit inside a car. SuperCard is given emphasis here because all of the recent major driver interface development projects (TravTek, Advance, this project) have used that application.

In selecting the hardware-software combination, it is important that the response time of the software be similar to the actual application. The author has noted that when less powerful Macintoshes are running these simulations (e.g., Mac IIcx with an accelerator board), system delays in responding to driver inputs can interfere with use of the interface. Macintosh Quadra 700-level computers and above provide adequate performance. Adequate performance is expected from all PowerPC computers running native code applications, but not 680x0 code in the emulation mode. Allowing simulations to be used in safety assessment tests is important to manufacturers, because those tests may occur early in the design process, not after a bug-free working model has been developed.

Having a vehicle to collect the data is not enough. Having the ability to analyze the data is equally important. Analyzing eye fixations and the other data can be done either manually or via software. Given that it takes 30 to 40 h to manually reduce each hour of eye fixation data, automation of this process is desired. Further, one is not interested in every moment of driving from a test route, only selected segments such as just before, during, and after an IVHS interface is used.

A particularly difficult task is identifying which segments of the test route are to be analyzed, and then filtering noise from the various desired signal segments prior to analysis. Data from the in-vehicle sensors of interest can be noisy.

## CONDUCTING THE TEST

Upon reporting to the test facility, the subject's driver license is checked (to see if it is current) and a standard vision test is performed. Subjects failing the State's vision requirements are excused.

The subject is then taken to the test vehicle and told to operate it as he or she normally would and use the product either on request or as needed. How that is achieved is product specific.

The amount of training given a subject prior to testing is a critical issue. While a few vehicles are now equipped with audio or videotapes to train drivers to operate new car systems, this is not the case for most vehicles. Further, when sold as used cars, it is unlikely those materials are provided to the second owner. Another mechanism for providing training is the demonstration provided by sales personnel to prospective buyers when shopping or on delivery. This cannot be assumed to be an especially effective means of training drivers because only the person that accepts delivery (not all drivers) receives instruction. While associated materials are provided for every new car (owner's manual), only some drivers read them.

If a system is truly easy to use, its operation should be intuitive and no training should be required. Ideally, no formal training should be allowed during on-the-road interface evaluations. If this does not match real use of these systems, then audio tape presentations of fixed durations (probably 5 m) should be allowed. In the ongoing project, subjects were asked to drive an 8-m training route that showed six route guidance screens, though no explanation of the navigation display was provided. However, any questions subjects asked were answered. Constraints on allowable training materials needs further consideration.

When training is completed, the subject should drive to the test route with the test system turned off. This time period is for the subject to become familiar with the test vehicle. The ideal situation would be to conduct the test using a car with which the subject is extremely familiar, namely their own car. However, it is not practical or cost effective to install the instrumentation required. Clearly, the longer a person drives a car, the more familiar they become with it. In contrast, rental cars typically are driven without any specific familiarization or instruction. The start of the interim test route proposed is approximately 20 to 30 m from Ford headquarters, and will take somewhat longer to reach from GM headquarters and Chrysler headquarters. In the on-the-road experiments conducted as part of this project, approximately 20 m was required to drive from Ann Arbor to the test site.[15] In addition, for one experiment, subjects drove an additional 5-m familiarization route (to become familiar with the navigation system) before the experiment began.[16] These routes seemed to provide subjects with an adequate knowledge of the handling of the vehicle. Subjects generally did not use the radio while driving and use of the climate control system was minimal. Sessions were conducted during the daytime in good weather, so neither the lights nor wiper were used. Further discussion of the issue of familiarity appears in the section on subject selection.

Route selection was determined by the existance of baseline data required by the protocol. As the road and traffic characteristics that affect driving become better understood, it should serve as an example of a route rather than the route.

Upon arriving at the test site, the interface of interest should be enabled and testing commence with the subject driving the test route. In the case of a navigation system, the subject needs to be told the destination and to enter it into the navigation computer. Other tasks may need to be created for other systems.

Immediately after completing the test route, the subject may be asked questions about the safety and ease of use of the interface (similar to those used in on-road tests as part of this project).[15,16] Eventually, responses to those questions may be added to the list of acceptance levels for the measures of interest.

Finally, the subject is paid and thanked for his or her time.

## TYPICAL INSTRUCTIONS TO TEST PARTICIPANTS

The instruction listed in this section assume the ideal case, that is, no training is required. These instructions are similar to those used in on-the-road tests of new driver interfaces conducted as part of this project.[15,16] In those tests written descriptions of the driver interfaces were not provided. In the assessment protocols described here, a draft owner's manual for the product driver interface should be in the test vehicle's glove compartment. The manual might be a separate document or incorporated into the vehicle owner's manual depending on the manufacturer's intentions. It is possible that for some products that manufacturers may provide videotapes, disks with multimedia presentations, CD-ROM's, or other media for demonstrating product use. If they are provided and it is found drivers use those media to learn about new interfaces, then some means for viewing the media should be provided.

### Navigation System/Traffic Information Example

"Here is a system to help you find your way. Tell it you are going to 123 Main Street. It will tell you how to get there. If you have any questions concerning operation of this device, you are on your own. There is an owner's manual in the glove compartment if you need it. While driving, please obey all traffic laws."

### Car Phone Example

"Here is a car phone. When you feel comfortable, please call my office at 764-4158 and talk to the secretary. She has some questions for you to answer while you drive. There is an owner's manual in the glove compartment if you need it. If you have any questions concerning operation of this device, you are on your own. While driving, please obey all traffic laws."

**IVSAWS/Vehicle Monitoring System Example**

"This car is equipped with an warning system you may not have seen before. Should warnings appear, take the appropriate action. If you have any questions concerning operation of this device, you are on your own. While driving, please obey all traffic laws. There is an owner's manual in the glove compartment if you need it."

**Information Retrieval System Example**

"This car is equipped with a system to provide information that may help you make your trips more efficient. At various times, you will be asked to retrieve various pieces of information and carry out other tasks. Do so at your own pace, being sure to always drive safely and obey all traffic laws. If you need to pull over to complete these tasks safely, feel free to do so. If you have any questions concerning operation of this device, you are on your own. There is an owner's manual in the glove compartment if you need it."

## DEPENDENT MEASURES

The number of potential performance measures in such studies is considerable. The table below shows the desired minimum sampling rate and accuracy for each of the measures to be recorded. These measures were selected based on their indicativeness, sensitivity to design differences, risk to drivers and experimenters, ease of measurement, ease of reduction and analysis, repeatability, acceptance by the scientific and engineering community, and the extent to which they fit into a desired experimental context. Details concerning the rationale for the selection of these measures is given in the previous report in this series.[17]

Table 2. Safety-related dependent measures.

| Dependent Measure | Accuracy | Sampling Frequency (Hz) |
|---|---|---|
| Speed | 0.16 km/h (0.1 mi/h) | 10 |
| Lateral Position | 33 cm (0.1 ft) | 10 |
| Steering Wheel Angle | 0.3 degree | 10 |
| Eye Gaze Direction | 1 degree | 30 |

From the above measures, mean speed, speed variance, average lane position, lane variance, steering wheel variance, and the frequency and duration of eye fixations to the road, other vehicles, mirrors, speedometer, in-vehicle display, and other locations as a group (e.g., the sky, the passenger) must be determined and reported. In the future it may also be desirable to have time-to-line crossing (TLC) and time to collision (TTC).[38,39] While TLC data have been collected by one European research team in real time, more experience with TLC is necessary before it becomes a requirement. TTC data have not yet been collected in real time on the road. While it is straightforward to describe how TTC can be measured, it too should be demonstrated before it becomes a requirement.

Particularly important are navigation errors, data which are required for tests of route guidance interfaces. Manufacturers are encouraged to collect more, direct, driver-interface- usability (operational) measures such as intervals between keypresses, control adjustment times, etc. Those measures tend to be interface specific. Usability measures should prove useful in diagnosing and eliminating product weaknesses.

## LEVELS OF ACCEPTANCE

To reiterate a critical point made earlier, it is very difficult to establish adequate levels of acceptance for most of these variables, because adequate baseline data for most of them are lacking. Accordingly, the values suggested in this section are given for the sole purpose of stimulating discussion of appropriate levels, not as required performance levels. Before any guidelines or assessment protocols are established, it is essential that baseline data be collected for a wide range of drivers, traffic conditions, and road geometries. Also essential are typical usage data for both conventional and IVHS driver interfaces.

Because drivers can make tradeoffs among various performance characteristics (and so, too, must designers) the author would argue against simply meeting minimum acceptance limits for all variables. In fact, in considering minimums, it was assumed that performance on other dimensions was good. For each measure of interest, three acceptance levels are given—best expected, desired/planned, and the worst case. One way to achieve the desired overall level of performance would be to require the average level across measures fall in the "desired/planned" category. Therefore, if performance for one measure only reached the worst-case level, performance should meet the "best expected" level for another measure to balance out. The tradeoff of measures is a topic needing further consideration.

### Speed Variance

Speed variances of 1.4 to 1.9 km/h (0.9 to 1.2 mi/h) on relatively straight expressways for a wide range of traffic densities, and 5.0 to 5.5 km/h (3.1 to 3.4 mi/h) on straight rural roads have been observed.[40] Noy reports a variance of 2.7 km/h (1.7 mi/h) for driving in a simulator in a car-following task on a curving road.[41] Standard deviations of 1.6 to 3.2 km/h (1 to 2 mi/h) have been reported as typical and standard deviations of 1.6 km/h (1 mi/h) have been observed in other researcher's simulator studies.[42]

It is unlikely drivers will be able to drive more steadily than cruise control units. Hence, the "best expected" level for speed variance can be determined from speed data for a variety of cars (with their cruise controls set) driven on a variety of roads with no interfering traffic. For the 1991 Honda Accord station wagon used in this project, data were collected on flat and slightly hilly expressways at 89 and 105 km/h (55 and 65 mi/h). There was one 2- to 3-m run for each combination with the cruise control set. The speed as a function of time was a triangular wave form with a peak-to-peak range of about 1.6 km/h (1 mi/h )with a period of

13 s. The standard deviation of speed ranged from 1.6 to 1.9 km/h (1.0 to 1.2 mi/h). The standard deviation was unaffected by hills or the set speed. More detail on speed variability when the cruise control is set may be available from automotive manufacturers or can be obtained experimentally (as was the case for the test vehicle) from a sample of vehicles. This data may be obtained by attaching a fifth wheel to the vehicle or utilizing the vehicle's own speed sensor (wheel pulse signal), as was done here. As research on intelligent cruise control is published, information on the performance of existing cruise control systems may be more widely disseminated.

To support the development of this assessment protocol, similar data were also collected for a single driver who was instructed to drive at 105 km/h (65 mi/h) on flat sections of an expressway with light traffic. Ignoring one outlier, the standard deviations for those runs were 1.4, 1.6, 1.6, 1.6, 2.1, 2.1, 2.2, 2.4, and 2.7 km/h (0.9, 1.0, 1.0, 1.0, 1.3, 1.3, 1.4, 1.5, and 1.7 mi/h). These data should not be viewed as a representative sample of the driving population as a whole, but simply as indicators of how well drivers can do when driving steadily is a high priority.

Also, considerable data on speed maintenance was collected.[15,16] Shown below is the distribution of the standard deviation of speeds for the baseline conditions [driving on straight sections of highway at 81, 89, and 105 km/h (50, 55, or 65 mi/h)].[16] Data shown in figure 2 are from 7 segments driven by 8 drivers (4 under age 30, 4 over age 65). Four of the data points are in excess of 3.6 km/h (2.25 mi/h) and only 1 exceeds 4 km/h (2.5 mi/h).

The mean standard deviation was approximately 2 km/h (1.25 mi/h). Notice the standard deviations are not normally distributed, an important point when considering the percentage of drivers exceeding some limit.

Given these data, for expressway driving on straight sections in light traffic, the acceptance levels that follow are suggested for the standard deviation of speed. Different values for curving roads are to be expected.

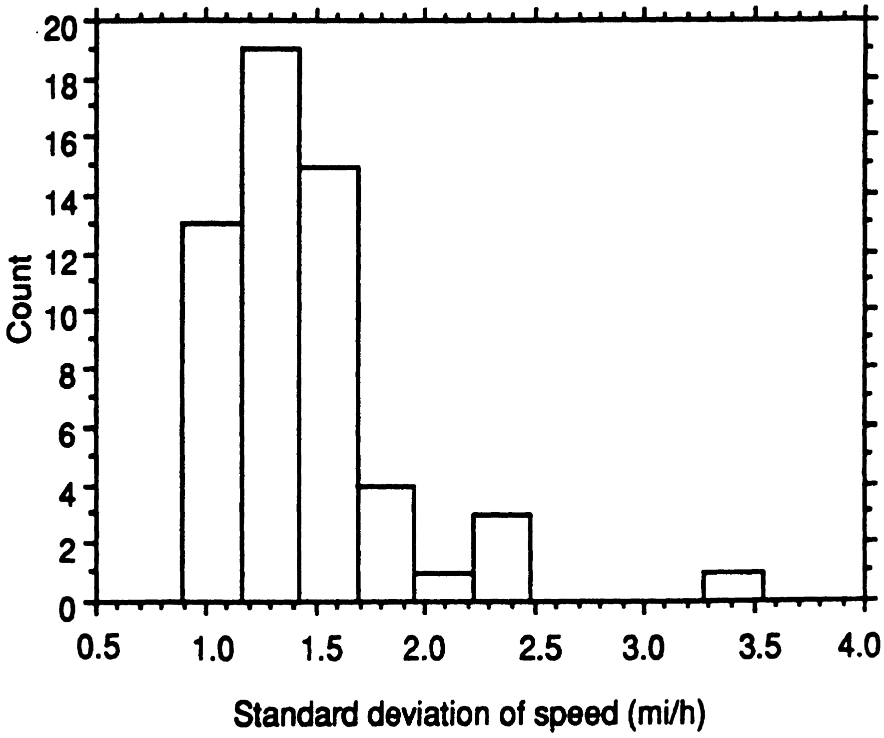


Figure 2. Distribution of standard deviation of speed.

Table 3. Standard deviation of speed.

| Acceptance Level | SD Speed (km/h, mi/h) |
|---|---|
| Best Expected | 1.6, 1.0<br>(comparable to cruise control) |
| Desired/Planned | 2.4, 1.5 |
| Worst Case | 3.2, 2.0 |

### Lateral Variability

With regard to lateral standard deviation, standard deviations have been reported of 0.39 to 0.59 ft for young drivers on an unused, four-lane road at roughly 13 and 81 km/h (50 mi/h).[43] Values of 21 cm (0.69 ft), 19 cm (0.62 ft), and 30 cm (0.98 ft), have been reported by three other groups.[44,45,56] Values of approximately 31 cm (1.0 ft) have been reported as being typical for driving curving roads.[47] These were all simulator-based studies with the age range being larger than in the other studies, except for one study, where half of the participants were 65 or older.[46,47] Values of 15 to 21 cm (0.5 to 0.7 ft) have been noted as typical.[41] Values of 16 to 31 cm (0.5 to 1.0 ft) have also been reported as typical.[48]

The assessment of lateral error has been examined with compensatory tracking (regulation against a random wind gust) and pursuit tracking (path following on a winding road) in a driving simulator based on data from six people.[49] Other details concerning the experiments have not been obtained. The following expressions for the lane-dispersion parameter in units of feet:

$$s_y = 0.17 + 0.012U_o + 0.321\ C_v \qquad \text{(compensatory)} \qquad (1)$$

$$s_y = -0.08 + 0.014U_o + 0.72\ C_v \qquad \text{(pursuit)} \qquad (2)$$

where $U_o$ = speed (ft/s)
$C_v$ = visibility parameter, approximately 0.1 for dashed lines and 300 ft visibility

Substituting 65 mi/h (95.3 ft/s) for the speed and 0.1 for the visibility parameter, the standard deviation is 41 cm (1.35 ft) for compensatory tasks and 45 c m (1.49 ft) for pursuit tasks. At 55 mi/h (80.7 ft/s) the values are 36 and 39 cm (1.17 and 1.28 ft), respectively.

A study examining the effect of alcohol on steering in a simulator showed steering errors of about 15 cm (0.5 ft) for a sample of drivers when sober and 31 cm (1.0 ft) when intoxicated (.10 BAC).[50]

It has been reported that the minimum lateral distance to the lane boundary was approximately 15 cm, just under 6 in.[51]

In this research project, data on driving performance were collected.[15,16] Figure 3 shows the lane standard deviations reported in reference 16.

To provide additional reference data, the instrumented test vehicle used for this project was driven on straight, level expressways at 105 km/h (65 mi/h) where drivers were instructed to "lightly" and "tightly" hold on to the wheel. There was some excess vibration due to a slight imbalance in one of the front tires but the vehicle was reasonably well aligned. Typically, the vehicle drifted about 46 cm (1.5 ft) every 6 s. Real vehicles are misaligned, but the distribution of misalignment in the population of vehicles on the road is unknown. To improve models of vehicle handling, data on this issue is desired.

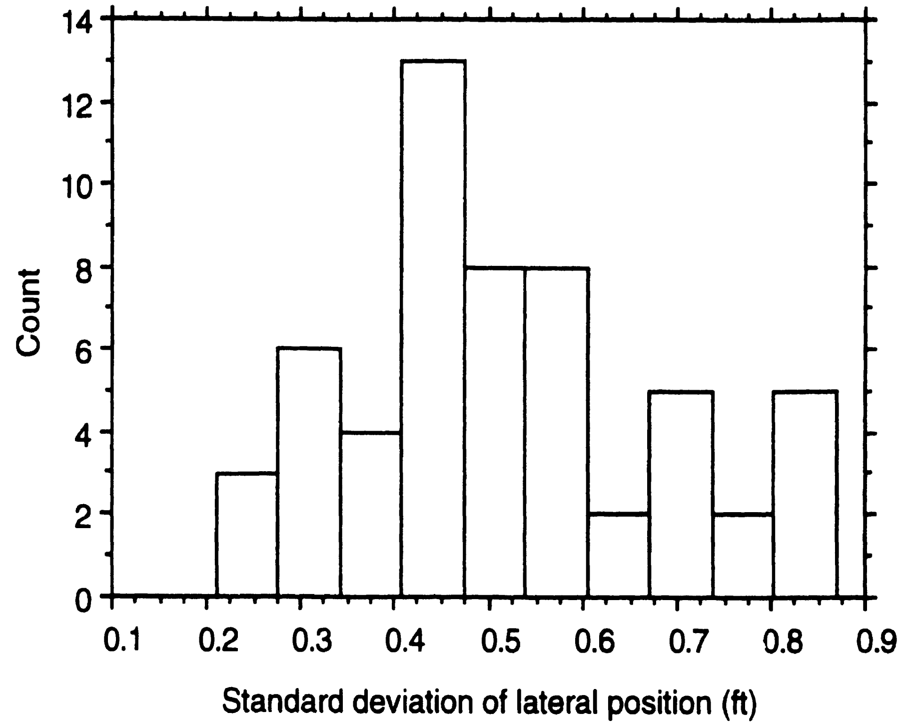


Figure 3. Distribution of lane standard deviations for the baseline condition.

What then should the acceptance levels be? Given the data from this project for driving on straight sections of expressways in light traffic and concurring reports from other sources, a standard deviation of 18 cm (0.6 ft) is the best one could expect. Lateral position is log-normal, not normally distributed, so using some multiple of standard deviations to reach a $p$ = 0.05 level is not appropriate. Using the values from figure 3, a standard deviation of 24 cm (0.8 ft) or less would include all but six of the cases explored. A value of 27 cm (0.9 ft) would include them all. The resulting limits are shown below.

Table 4. Standard deviation of lateral position.

| Acceptance Level | SD Lateral Position (cm, ft) |
|---|---|
| Best Expected | 18, 0.6 |
| Desired/Planned | 24, 0.8 |
| Worst Case | 27, 0.9 |

To put lane variance in context, a typical expressway lane is 3.7 m (12 ft) wide and a large car about 1.8 m (6 ft) wide, leaving 0.92 m (3 ft) on either side of the car. If the lateral position was normally distributed (which it is not), then a standard deviation of 46 cm (1.5 ft) would result in encroachment on the lane edge approximately 4 percent of the time, far too often. Moreover, the standard deviation of lateral position changes considerably with the number and size of curves, which are generally not specified.

## Eye Fixations

With regard to eye fixations, there are two sets of data that suggest performance limits for situations where drivers are asked to complete a specific task over a short period of time such as identifying the next cross street. Shown in figure 4 are the data from reference 52, and in figure 5, the data from reference 53. The desired/planned level should correspond to the outer group of commonly used controls (an average of 2.5 glances of 1.1 s or less) since navigation systems are new and drivers have little experience with them. Driver reactions suggest that using the cassette tape unit (4 glances) and establishing the destination direction (1.2 s/glance), were too difficult, resulting in the eye-fixation recommendations that follow. These recommendations are for navigation tasks listed in figure 5 and the most common task, the name or number of the street or road for the next turn. Note the close relationship between Zwahlen's recommendations and the author's interpretation of Wierwille's data that follows.[51,52]

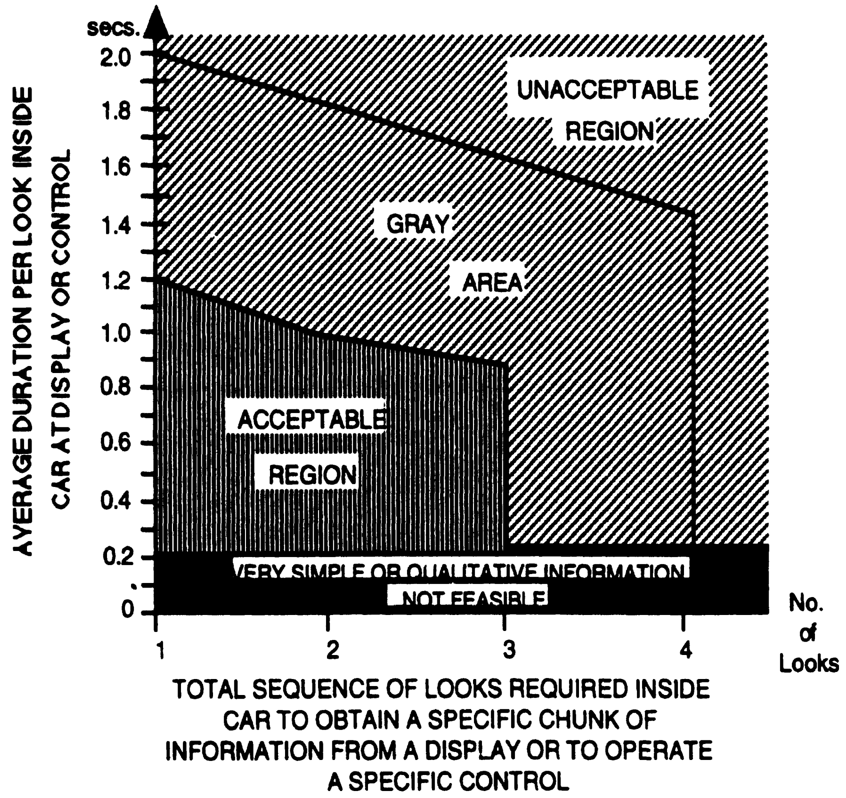


Figure 4. Eye fixation recommendations.[51]

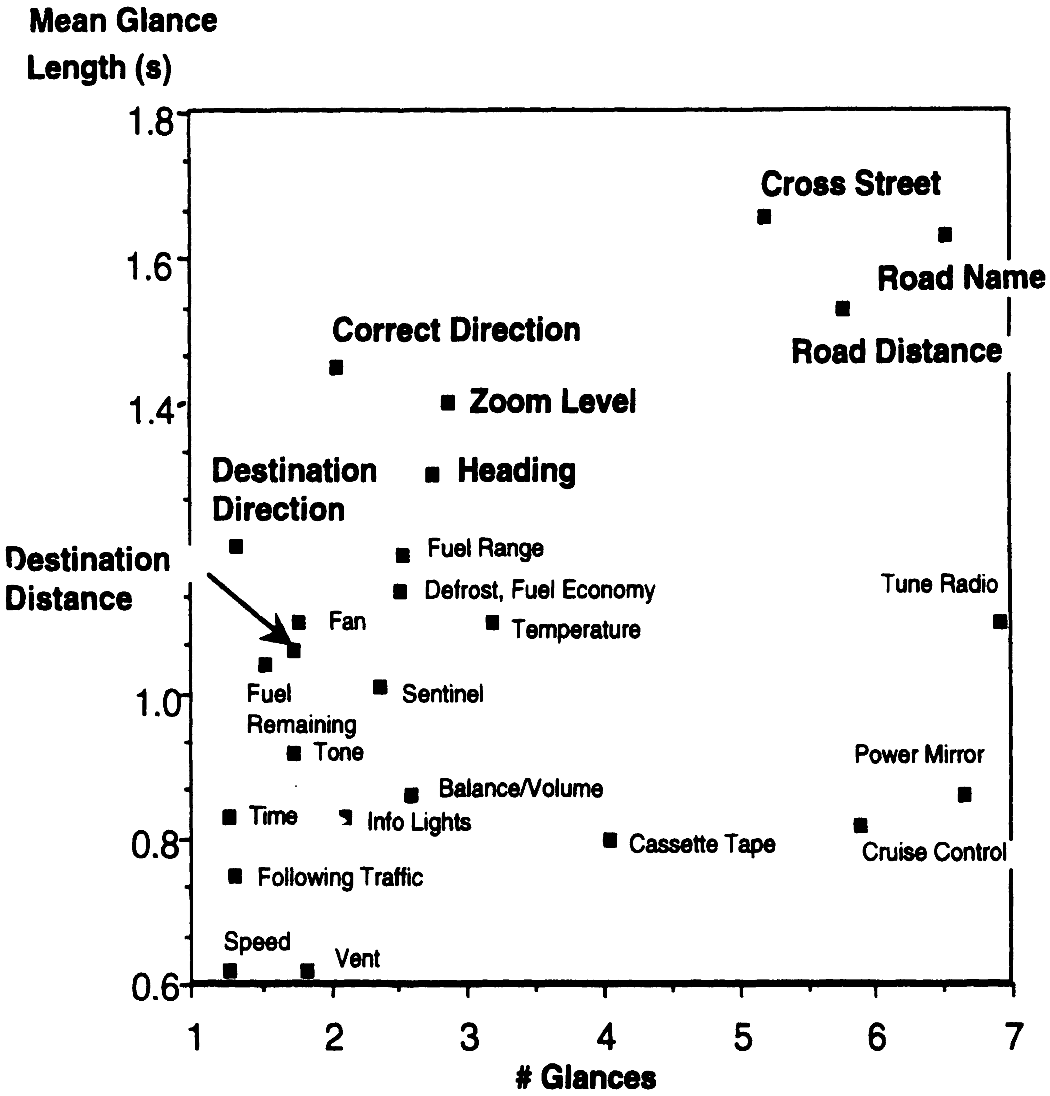


Figure 5. Number of glances versus glance duration.[47]

Table 5. Glance limits.

| Acceptance Level | # of Glances | Glance Duration (s) |
|---|---|---|
| Best Expected | 2.0 | 0.9 |
| Desired/Planned | 2.5 | 1.1 |
| Worst Case | 4.0 | 1.2 |

Readers should note that the criteria used for setting performance levels for eye fixations are different from those of the other performance variables. The speed and lane-variance criteria primarily consider normal behavior; data that are lacking for eye fixations. Hence, the eye fixation limits were based on ease of use. If these systems are to be accepted by drivers, they must be easy to use. What little information that does exist relating to the number of eye fixations required suggests that it is possible to drive (on a closed road without traffic) and look at the road very infrequently.[54] It is unknown what the requirements are for safe driving when traffic is present, the typical case.

Finally, recent experience has shown that for driver interfaces that use HUD's as the display, it may be difficult to distinguish looking at the road with looking at the HUD.[14] For auditory systems, where there is no visual display, the need for eye-fixation data is uncertain.

The major difficulty with the eye-fixation-data acceptance levels described previously is that it is assumed the task of using a navigation system is discrete, which it is not (except in experiments). These fixation limitations may also be appropriate for specific screens that are retrieved by drivers ("yellow pages") or screens not used in an ongoing manner (traffic information). When using a real or simulated navigation system, drivers continually refer to the display.[14,15,16] The test route used in this project consisted of a variety of expressways, main roads, and residential streets. For the 19-turn route turns were 30 s to 4 m apart (mean of 70 s).[15] For that route, there were about 7 glances per minute for younger drivers and 8/minute for older drivers or approximately 1 glance every 8 s. The glance frequency ranged from 1.5 to 12.9/minute for that sample. These values seem a bit high, but readers should recall that turns were rather frequent on the test route. For younger drivers there was a mean 9.2 glances to the navigation display per turn. For older drivers, there were 11.4 glances. Given the lack of understanding of glance behavior, it is premature to speculate as to what number of glances correspond to best expected, desired/planned, and worst case for the test route.

If ease of use is to be seriously considered, there should be performance measures for actually using each system; measures that are system specific. The most important of those is navigation, and for navigation the key measure is the number of turn errors. A turn error occurs when the driver fails to make a requested turn, turns the wrong way, or turns when not requested to do so. In one study, people drove routes involving 6, 7, or 13 turns (average of 8.66 turns).[55] There was an average of 1.12 errors (12.9 percent) for the auditory system, 1.87 (21.6 percent) for the map, and 1.64 (18.9 percent) when both the map and the auditory system were used. Schraagen had people drive four routes in unfamiliar cities using either a photocopy of a map or a photocopy with colored dots highlighting turns and showing signs to look for at turn points.[56] There was a total of 20 turns. Schraagen reports the average number of errors across the four routes was 1.6 for men and 2.4 for women, suggesting that the turn-error rate was 10 percent.

In the research conducted as part of this project there were two classes of errors, execution errors (situations where drivers missed a turn or turned early) and near errors (where they said they were confused or almost made an execution error). In the initial on-the-road evaluation there were 25 total errors (near errors plus execution errors) made by the 30 drivers on a route with 19 turns.[15] Of those errors, 10 were execution errors and 15 were near errors. In the second on-the-road evaluation, using a slightly modified route, there were 9 total errors, 3 of which were execution errors.[16] Across these two experiments the total error rate was 4.6 percent, (1.8 percent execution errors, 2.8 percent near errors). For this particular route, looking at the types of errors made by drivers and considering comments from the experimenters, the author believes error rates could be reduced by as much as 50 percent if the driver interface was optimized. Further, it is believed that execution errors that are double those from the research conducted as part of this project are minimally acceptable, though supporting experimental evidence is lacking. The levels are considerably below other values appearing in the literature. It may be that better performance was achieved because the route was easier to drive. Again, readers should bear in mind that the primary purpose of suggesting levels is to encourage discussion and exploration of acceptable levels, not to offer legal limits for certification.

Based on these data and the author's experience, the levels that follow seem reasonable for the route examined. As a test of reasonableness, the last experiment conducted as part of this project can be thought of as validating the protocol. For that eight-driver sample (a very limited one) there were three execution errors, leading to an execution error rate of 2.0 percent, approaching the desired/planned acceptance level.

Table 6. Turn error limits.

| Acceptance Level | % of Turn Errors (Execution) |
|---|---|
| Best Expected | 0.9 |
| Desired/Planned | 1.8 |
| Worst Case | 3.6 |

Note: The percentage of turn errors is the number of wrong turns the driver makes divided by the number of turns the driver should make. Most turn errors are drivers turning where they should not, or missing turns, rather than turning the wrong way.

## WHERE SHOULD TESTS BE CONDUCTED?

Of the many test requirements, this is certain to be one of the most controversial because of the problems it will pose to organizations without a facility in the state of Michigan. The test route or route should have (1) documented driver performance data for driving the route with the information systems of interest, (2) contain a variety of intersections and road types, (3) be conveniently located for manufacturers whose products will be assessed, and (4) be accessible throughout the year and for many hours each day. A key issue is whether there should be one route or multiple routes. This depends on whether test conditions can be specified sufficiently such that results can be replicated (and conducted efficiently) at multiple sites. On-road performance in the use of in-vehicle systems is affected by many factors: traffic volume, time of day, weather (clear, rain, snow, fog, etc.), the type of road (expressway, city street, etc.), the distance between turns, the number of various types of intersections (signalized, signed, etc.), lane width, speed limits, curve radii, sight distances, shoulder widths, etc. Some, but not all, of these factors are captured in Wierwille's workload model.[57] This model has not been independently verified. Since the effect of many of these factors is unknown, all of them must be constrained so test results are consistent and reliable. Because there are so many factors to specify and because finding routes matching all characteristics will be extremely difficult, and because the influence of many factors is unknown, it makes more sense to pick a known route for this initial test protocol.

The only route for which all of the desired driver performance data exist is the route used in recent tests conducted as part of this project (criterion 1). That route contains a wide variety of intersections and road types (criterion 2). As noted earlier, the proposed approach assumes the car manufacturers will conduct assessment tests, so the site should be convenient for them. Manufacturers were chosen because of the availability of people, equipment, and facilities, adequate checks against falsifying the results, and the opportunity for assessments to lead to safer and easier to use products.

All of the U.S. car manufacturers have their main engineering offices in southeastern Michigan (General Motors in Warren, Ford in Dearborn, Chrysler in Auburn Hills). Development of navigation systems is most likely to occur in their main engineering offices, though some of the General Motors efforts will occur at their Delco facilities in Kokomo, Indiana. Significant development work is unlikely to occur in southern California (where aerospace and defense contractors are located) or near the Federal Government in Washington, DC, two other suggested locations for the test route. To work with the major U.S. automotive manufacturers, all of the major automotive suppliers also have offices in southeastern Michigan (usually in Troy or Southfield). Most foreign manufacturers have offices in southeastern Michigan, and several have offices in Ann Arbor (specifically Toyota, Mazda, Nissan, Honda, Mitsubishi Motors), because the Environmental Protection Agency (EPA) emissions assessment laboratory is there. Foreign manufacturers are most likely to do the primary engineering work on new information systems in their home countries. Because it is convenient to the manufacturers, DOT has occasionally held meetings at the Ann Arbor EPA lab relating to safety activities. For those reasons, it would be prudent to have the test route located in southeastern Michigan, somewhere between Ann Arbor, Warren, Auburn Hills, and Dearborn (criterion 3).

On-road testing is not without its drawbacks. The road network can change over a period of years. New roads are added and others are upgraded or widened. Businesses come and go, altering the roadside clutter. Finding an area that is relatively stable is important, especially with regard to traffic. The test route suggested does not suffer from congestion. However, in Michigan, during the spring and fall, experimenters need to contend with rain, and in the winter, from November to February or sometimes March, snow can interfere with testing. Thus, the test route is not ideal with regard to criterion 4, but this criterion may be less important than the other criteria.

Accordingly, to satisfy immediate needs, the author recommends using the Belleville to Canton, Michigan route chosen for the on-road tests with the revised sequence being preferred because it contains more long, straight sections useful for calibration and collecting baseline data (driving only, no information systems used). (The initial route is described in reference 15. The revised route is described in reference 16.) These straight sections are needed for lane variance measurements. The route has a wide variety of intersection types, though it lacks traffic circles. Testing should be conducted under conditions when traffic is free-flow (9 a.m. to noon, 1 p.m. to 3:30 p.m.) in clear weather during the daytime only. Nighttime testing should be added to the protocol in the future because the lighting characteristics of visual displays (contrast levels and directions, colors used, etc.) can be very different at night. To date, none of the experiments from which the acceptance criteria were developed involve night driving, nor do most of the on-road tests of IVHS driver interfaces. Hence, specifying how nighttime conditions might differ is not possible given the absence of material in the literature. Night driving is important considering the relatively greater risk of accidents during that time and the absence of environmental clues (easy-to-read signs) on which the driver can rely for route guidance.

The author would suggest that, in the future, the route be modified to eliminate sections that are in poor repair. There also has been some thought that the route might be too long and limited in the variety of turn decisions to be made. It also lacks driving in a downtown section, a commercial (strip mall) district, and a segment where there is a long period between turns. Finally, as more is learned about how road geometry, traffic, weather, and other factors associated with a route influence driving performance and behavior, other test sites need to be considered.

In some situations, the current route may be too long. For some eye-fixation-recording systems, drivers cannot comfortably wear the head and eye sensors for the entire trip. That will obviously depend upon the type of eye fixation recording system a particular vendor may choose for assessments. Specifying a new abbreviated route (and collecting baseline data for it) is beyond the scope of the project. For details of the existing route, see the reports for this project describing the on-road testing.[15,16]

An added advantage of using a specific route is that best, desired, and worst case levels for the performance measures can readily be specified as has been done in this document.

## INDEPENDENT MEASURES

In specifying the route, time of day (daytime, non-rush hour), and weather conditions (clear), few independent variables remain. Since the purpose of this test is to determine whether various systems are safe, rather than to explore design alternatives, it is desirable to minimize the number of test conditions. As was noted earlier, the procedure should eventually be modified to include nighttime driving.

Though it is not a requirement for assessment tests, manufacturers may want to add in their own tests (which might include testing several interfaces designs, but only reporting the results for the version of the system they intend to produce.

## SUBJECTS

Subjects in these experiments should be licensed drivers who meet the vision requirements for driving, and who are in good health.

Differences due to both sex and age are of interest. In typical experiments examining user interfaces, at least two and sometimes three age groups are tested. Ignoring novice drivers, driver age seems to have a small effect on performance up to age 50, though there are small differences between groups above and below age 30. (This, of course, ignores accident rates, which drop precipitously after age 25.) Clearly, older drivers (over age 65) do significantly worse. There is a developing trend to treat older drivers as consisting of two separate groups—66 to 75, and over 75. Ideally, one should test only the group that has the most difficulty with these systems, those over 75. However, the acceptance limits given earlier are for all drivers; hence, a mixture of older and younger drivers should be tested.

An alternative strategy is to select drivers based on a market segment analysis. For example, if a product is designed for a high-end sedan (e.g., Lincoln Continental or Cadillac Coupe DeVille), the sample could be restricted to owners of those and competitive cars. Drivers of those vehicles could be identified from lists provided by the R.L. Polk service, though there may be other sources. Unfortunately, acceptance data do not exist for particular market segments. Further, interfaces are likely to be adopted by other car lines in the future, requiring re-testing with additional groups of drivers fitting the revised demographics.

As mentioned earlier, subject familiarity with the interface being tested or similar interfaces is an issue. At the present time, all drivers will have had some experience with vehicle monitoring systems as they are in all cars, some drivers will have had experience with car phones, very few drivers have had experience with navigation and traffic information systems, and only subjects in two experiments conducted as part of this project will have had experience with IVSAWS (because that system had not been installed in any production vehicles). Hence, at the present time, the typical case (and worst case) is no experience with the interface in question. Accordingly, subjects should be selected on that basis. A potential exception concerns tests of car phones. The fraction of buyers of car phones that have prior experience in using them is unknown. That information, when it becomes available, should be used to target subject selection. Over the next 10 years as the market penetration of new driver interfaces increases, familiarity criteria for subject selection should be reviewed.

To calculate the number of subjects required, a power analysis could be computed for the dependent measures of interest. A power analysis requires estimates of age and other factor-related variances, which at present are imprecise, so that approach is not a good choice. An alternative is to consider what constitutes a manageable experiment that meets the constraints of a product schedule. The author's experience has been that, when more than 40 drivers are to be tested in an experiment, recruiting becomes much more complicated. If gender and age (two groups, under 30 and over 65) are to be explored, there would be 10 drivers per cell, a reasonable number. Given the likely duration of a test run, the need to prepare a vehicle before each run, maintain it, and download data, two subjects per day is the best one can expect, especially if, to be consistent, a single experimenter is to be used.

At that rate, it will require 20 working days (4 weeks) to collect the driver data and at least double that time (8 weeks) to plan the test and prepare the vehicle (assuming one is on hand). A similar time period (another 8 weeks) is necessary to analyze the results and prepare a report (assuming the evaluation team can focus only on that report). Thus, a 40-subject test program could take 20 weeks, far too long. New car development times vary quite widely, from 6 years for the Ford Mondeo to 36 months for the Chrysler Viper. Durations are 36 to 42 months are becoming common, and there is pressure to reduce that time further. Because of the time required for tooling for the interface and other pre-production testing, requiring 20 weeks for testing would push specification of the interface well into the early phases of design, well before the vehicle concept is adequately defined.

With careful subject selection, it may be possible to reduce the number of subjects to 32 (8 per cell), shaving 4 days off of the subject testing and 8 days off of the analysis and report writing phase, for a total of 17 weeks. More thought than is possible within the domain of the project must be given to the planning and execution of these assessment tests, because test duration and cost could have negative impact on the affordability and speedy implementation of new driver interfaces. The cost of identifying safety problems and enhancing usability must be balanced with the risk of potential hazards and the customer-perceived cost of diminished usability.

Subjects are generally recruited from lists on hand and via newspaper ads. Recruiting should be from a demographically broad pool. For example, selecting subjects from only groups of automotive engineers or college students is inappropriate. For navigation and other systems that have little market penetration, test subjects should have had no prior experience with them. It is unclear what level of prior experience with car phones is appropriate.

In Ann Arbor, payment to participants is typically $12/h for daytime on-the-road time and the pretest screening of visual acuity. For nighttime work, $15/h is typical.

## TEST DURATION

During the initial on-the-road experiment conducted as part of this project, each session took 2.5 h/driver to complete. This included signing the consent form, obtaining biographical data and vision screening, calibrating the eye camera and checking the equipment, driving from Ann Arbor to Belleville (about 25 min), setting up the eye camera (10 min) driving the test route from Belleville to Canton (35 min), responding to a post-test questionnaire (10 min), driving back to Ann Arbor (about 25 min), and debriefing the subject. The time it will take others to test each driver will vary. The travel time from Dearborn (Ford) to the start and end points of the route is approximately the same as from Ann Arbor. The time from Warren (GM) and Auburn Hills (Chrysler) is somewhat longer. The test duration will also depend on how much debriefing the experimenters do (to identify product improvements) and other tasks that may be piggybacked on the experiment. While 2.5 h is a lengthy test session, drivers have not had problems completing the experiment because they do not drive for the entire session. The author is concerned that if the driving task was extended considerably, problems with fatigue might occur. One way to avoid such problems would be to have the experimenter drive the return trip to Warren and Auburn Hills.

Collecting all of the data in a single session has significant scheduling advantages. If the testing organization was asked to carry comparisons of alternative interfaces, then multiple sessions would be needed. There might be some learning transfer across systems, and learning should not be ignored when interpreting assessment results.

## DATA ANALYSIS/SCORING

Data reduction in field studies is a major concern. The first step involves transfer from the vehicle (often on a cartridge tape or a Bernoulli cartridge) to digital media in the lab. Elimination of spurious data is often necessary. It is done on a subject-by-subject basis, looking at histograms and statistical summaries of each dependent measure to identify outliers and other problem data points. Lane tracking and speed signals can be noisy and need to be filtered.

The second step involves generating an initial data summary. For example, in many situations, the raw-eye-fixation data is a videotape of a subject's face. Conversion of that information to a list of objects fixated takes 40 h for each hour of data collected. Where data collection is automated, reduction of the data involves segmenting the trip and eliminating those segments not of interest.

Concurrently, synchronization of the data streams (video, audio, computer recorded) must occur. This may involve playback of multiple VCR's at once, combining data files, or the use of Society of Motion Picture and Television Engineers (SMPTE) time code.

The last step involves computing tests of significance, often using Analysis of Variance (ANOVA), though sometimes Analysis of Covariance (ANCOVA). Because the data sets can be massive, reduction of data from these studies can be extremely time consuming. Some analyses can take several hours of personal computer time for each subject.

This section only briefly describes data analysis. The cost and time required are considerable, discouraging scientists and engineers from collecting on-road data.

Readers interested in experiments following this protocol should examine references 15 and 16.

# DRIVER-LICENSING-OFFICE TESTS

## PURPOSE/ISSUE AND SYSTEMS TARGETED

The purpose of the tests conducted in driver licensing offices is to determine whether drivers have a basic understanding of information presented to them. For the purposes of certifying the safety and usability of a system, this test will primarily be used for enhanced systems for which previous versions of the interface have been assessed using on-road tests. Those enhancements might include minor changes in the arrangement of graphics, the addition or deletion of color, changes in wording or symbols, etc. This procedure is most important for systems in which the view of the road ahead at any given moment does not provide contextual cues that are intimately linked to the system. For example, the protocol should be fine for vehicle monitoring systems where a verbal description of the driving scenario should be sufficient.[13,58] Similarly, cellular phone interfaces can be evaluated without strong contextual cues. However, for navigation, traffic information, and road hazard warning systems, the applicability of surveys for evaluating interfaces depends on the question. For example, in one of the experiments conducted as part of this project, alternative formats for indicating the location of hazards were examined.[16] Drivers sat at an intersection in a real car and were shown cards (placed where they might be in a real system) with warnings on them and asked to point out the window to the hazard. While some of the contextual cues could have been provided in an office study via a panoramic photograph, it is not clear if that would be sufficient. Similarly, there are times when context can be provided for traffic information and navigation displays using pictures.

As an example of the limits of paper and pencil methods, drivers were shown plan, aerial, and perspective views of intersections in an experiment conducted as part of this project.[9] Drivers were asked to interpret and rank them. Differences in interpretation errors between designs were small. In a subsequent experiment, response times to navigation displays differing in the view of intersections were collected . In that experiment, there were significant differences in both response times and errors.

The author expects that, over time, the majority of the testing will be paper and pencil evaluations, though there is no reason why testing organizations would not be allowed to use part-task simulations instead should they wish to do so. When systems are initially produced, they will be subjected to the expensive and time-consuming tests described in the previous section. Thereafter, most of the changes will be slight modifications or enhancements and the type of testing described here will occur. While the individual changes made for various model years may be small relative to the previous version, the cumulative effect of these changes may result in a design quite different from that last tested on the road. In that case, on-road retesting is necessary.

(Readers interested in examples of tests of this type should see references 10, 12, 16, 51, and 59.)

## HOW THE TEST IS CONDUCTED

Participants are shown individually a static mockup of the display and a sketch showing where it will be placed in a vehicle. While these tests could be conducted using rapid prototypes of the interface (running on a laptop computer), these tests may often be performed early in the development cycle, when interface storyboards are available, not prototypes. Should prototypes exist (usually derived by modifying the prototype of a previous version), they should be used.

The participant's task is to say what each element in the display means. The easiest way to record performance is for the experimenter to take notes on what the subject says, trying to capture quotes verbatim, except for minor speech errors (ums, uhs, wells, etc.). Laptop computers are an alternative to handwritten notes because they tend to be more complete, but cannot be used for subjects standing in line. The drawback of laptop computers is that they can be intimidating. When their handwriting recognition accuracy improves in the future, personal data assistants (PDA) may be the preferred recording device. Data can also be recorded using a voice-activated tape recorder and later transcribed, though background noise will degrade the recording quality. The nature of the conversation with participants may not be conducive to audio taping. While taping provides more accuracy, quick turnaround is generally desired, and the incremental accuracy is of little benefit.

After the test is completed, the experimenter will record the subject's sex, age, occupation, and other pertinent biographical data.

While testing can be done when standing in line, many offices (e.g., those in Michigan) have people take numbers to hold their places, so going to a card table does not cause one to lose a place in line. In other instances, the participant's partner (if accompanied by one) can hold his or her place. Waiting times (in line) of 10 to 15 min are desired to allow adequate time for interviews.

## TYPICAL INSTRUCTIONS

“Here is a display that might appear in a car of the future. Here is where it will be in the car. Tell me what you think this display is telling you and what each item is for." If the subject does not explain part of the display then ask, “What is this telling you?” Also ask, “Why do you think that?"

## DEPENDENT MEASURES

The time required for a subject to begin saying something substantive (the time after “gee, let me see” or "let me think about this for a moment") is a potential measure. The rationale is that if a visual display cannot be interpreted quickly, fixations to it will be long and numerous, drawing the driver's attention from the road, thus compromising driver safety. Obviously a display that cannot be interpreted quickly is more difficult to use than a display whose interpretation is immediate. The author believes that responses should begin within 10 s of the time the initial instructions are completed. If it takes longer (and usually it is much longer, approximately a minute), the display is not readily understood and fails the test for that individual.

The connection between minimizing errors in reading the display and ease of use should be transparent. In terms of safety, misinterpreting the display could lead to drivers turning on to and driving on one-way streets in the wrong driving, driving on streets that have high accident rates, or simply driving more (to correct errors), hence increasing their exposure to accident situations.

For a good design, 80 percent of the responses should be correct and within the time deadline. The 80 percent value has not been validated and is based on the author's judgment. There are no data concerning verbal response times in informal settings in the literature. To put this value in perspective, for studies of symbols, 75 percent correct (with no time constraint) has become the standard of acceptance.[18] In a free-response naming experiment, the horn symbol was identified correctly 97 percent of the time, the battery 94 percent, the turn signal 91 percent, the trunk release 84 percent, the fuel symbol 82 percent and the high-beam symbol 80 percent.[60] In some sense, the rate that is appropriate will depend upon what the questions are and how they are asked. For example, in the Saunby, et al research, there were large practical differences in the scores for free response (what does this symbol represent) and matching tasks (which of these is the correct name for this symbol).[53]

In the highway literature, the critical value (for speed levels, sight distances, etc.) is the 85 percentile. The author does not know the rationale for selecting that value.

Considering the above, the following desired performance levels are suggested.

Table 7. Percent error limits for driver-licensing-office studies.

| Acceptance Level | % Correct |
|---|---|
| Best Expected | 90 |
| Desired/Planned | 80 |
| Worst Case | 75 |

How participants respond is more indicative of their understanding than the time taken. Answers such as I don't know, blank stares, or similar responses are obvious indications of a failure to understand. Knowledge of the logic behind wrong answers may suggest opportunities for redesign. Again, this information is not required for determining minimum safety levels, but is useful for improving design.

Finally, before an organization uses this method in place of on-road testing, it needs to present comparable, baseline data for a system evaluated both on the road and in driver licensing offices to calibrate the acceptance levels.

## INDEPENDENT MEASURES

Often the purpose of research is to evaluate a single display rather than to compare alternatives. Where alternatives are compared, independent variables often include the use of symbols or abbreviations as labels, various abbreviations, alternative arrangements of graphic elements, or graphic formats.

## TEST DURATION

It is very important that the test be completed while the subject is waiting to be served, a time period that varies from office to office and day to day. Field studies of this type are targeted to be completed within 10 min because people are generally willing to devote that much time, though sometimes 15 min is acceptable. In planning these studies, experimenters should realize that 2 to 3 min are devoted to administrative overhead (explaining what the experiment is about, giving instructions, obtaining biographical information, etc.). Signing a consent form can be a major delay. Because of the innocuous nature of these experiments, consent forms are not necessary.

## EQUIPMENT, MATERIALS, SOFTWARE

Requirements for these items are minimal, so that tests can be conducted quickly and inexpensively. A static mockup of the display is typically mounted on foam core or cardboard. The display image usually is produced by a laser printer. Colors should be indicative (e.g., a green light should look green) but color tolerances are unimportant. Coloring can be done by hand. The image should be laminated to avoid degradation due to handling.

As was noted earlier, data can be recorded using any laptop computer running a word processing program with the response form already keyed in. The data can also be recorded using paper forms on a clipboard. For experimenter and subject comfort, a card table and chairs may be needed.

## PARTICIPANTS

Licensed drivers should be the test subjects. The ideal site for this type of test is a driver-licensing office. At most offices there is a predictable stream of arriving drivers (sometimes in pairs) who otherwise have nothing to do while waiting in line, especially at peak times. Licensing offices usually offer a good demographic sample of the area. There tends to be a good balance of men and women, and of various age groups. Some thought needs to be given to the site. For example, the Beverly Hills, California office would probably not give a very broad demographic sample of adults. Subjects are generally not paid, though the offer of a candy bar or similar snack helps encourage participation.

Paid subjects may also be tested at shopping malls, or recruited via newspaper ads to visit an experimenter's facility.

Prior to testing, permission must be obtained from the driver-licensing bureau in the State capital to conduct tests. This generally requires a phone call followed by a letter (possibly with a draft survey). The State licensing bureau should make the initial contact with the local office desired for the test site. The time required for logistics to be completed will vary from State to State. In Michigan, a minimum of 2 weeks should be allowed for lead time. For this project, there was no difficulty obtaining permission to conduct studies, in part because of the contractor's long working relationship with the licensing office, its emphasis on safety (a mutual concern), and its status as a State institution. Ford has gotten permission to carry out similar studies. Private companies conducting reasonable, brief, mandated, safety surveys should not have difficulty obtaining approval.

The larger the sample size, the greater the reliability of the result (and the greater the cost). In the author's opinion, a sample of 100 drivers should be sufficient. At a reasonably busy office, one can expect to get 20 drivers/day (though the number will vary widely from office to office and day to day). In an office with peak busy periods, five to ten drivers per day is more likely.

## TEST CONDITIONS

Driver-licensing offices are generally very busy places with many distractions and little privacy. Ideally, subjects should be recruited as they arrive at the licensing office. Because people may sit in rows of chairs in a waiting area, subjects and potential subjects may hear or see the interactions of others. Having a card table off to the side avoids such problems, though it adds the effort of leaving one's seat, which some are reluctant to do.

## DATA ANALYSIS/SCORING

Data analysis usually takes the form of scoring responses as correct, partially correct, or incorrect. It is very important to develop scoring rules before the data are collected since personal biases will otherwise affect the outcome of an experiment. This will require some brainstorming by the research team, and possibly some pilot data. For example, when drivers are shown the ISO symbol for high beam (the bullet), responses include "lights," "exterior lights," "headlights," "main lights," "driving lights," "brights," etc., which are correct to varying degrees. Because subjects will offer unanticipated responses, some minor changes to the scoring scheme may be required after the data is collected.

Finally, while not a requirement, it is extremely valuable to record exactly what subjects said (and sometimes, how they said it), especially where designers are not involved in the actual testing. In a study evaluating a vehicle maintenance monitor, a hypothetical subject was videotaped interacting with the experimenter.[49] Because it can interfere with licensing-office operations, taping all subjects is not recommended. However, a recording of just one typical subject, along with the data from the others, is effective in convincing designers that changes are needed. In many organizations, the major problems are not the engineering details of the interface, but getting people to appreciate that the human factors problems are real and significant, and are not something users will readily overcome with training or adaptation. A good protocol should not only verify that an interface is safe and easy to use, but foster development of those types of interfaces.

## OTHER TEST OPTIONS

There are several other options for conducting tests to assess the safety and ease of use of IVHS systems. Tests on race tracks, proving grounds, and closed driving courses are options where risk is an issue. More likely options for tests related to driver information systems include laboratory response-time tasks, part task simulations using videotaped road segments or Motion Picture Experts Group (MPEG) compressed segments stored on hard disk, and driving simulator tests. These methods can be extremely useful in comparing design options and many have been explored in this research project. However, for most of them, acceptance criteria are lacking. In addition, where criteria do exist, linking them to on-road performance is difficult. None of these methods is currently recommended for use as a formal certification procedure, though that could change in the future.

### RESPONSE TIME (RT) TASKS

Following are brief descriptions of how these methods might be applied. In response- time tasks, participants see or hear information and respond by pressing a key or saying something. Timing is the most clear when the stimulus is visual and the response is a button press, because all of the information is available at the onset of the stimulus and the complete response occurs in an instant. Dependent measures include response time (to the nearest millisecond) and error rate, with more significance attached to response time because it is a more stable measure.

As an example, in a recent study, drivers seated in a mockup of a car were shown a slide of an intersection and a slide of an in-vehicle orientation display, simultaneously.[8] Drivers pressed buttons to indicate whether the scene and the slide were of the same or different intersections. Where they differed, the geometry was mismatched (e.g., the road scene showed a T and the in-vehicle display showed a Y). This experiment was used to compare a head-up display (HUD) with an instrument panel (IP) display location, the view of the intersection desired (plan versus aerial versus perspective), and whether roads should be outline or solid.

The questions addressed, and the task the driver executes, are specific to the information subsystem of interest. Tables 8 through 10 show examples for the navigation, In-Vehicle Safety and Advisory Warning System (IVSAWS), and vehicle monitoring system driver interfaces.

Table 8. Navigation questions.

| **Question** | **Voice Response** | **Button Press** |
|---|---|---|
| Should you turn/exit now? | yes/no | (y/n) |
| What road are you on? | name it | x |
| What is the next cross street or exit? | name it | x |
| What is the name of the street of exit for the next maneuver? | name it | x |
| How far is it? | give # | keypad |
| What lane should you be in? | left/right | l/r |
| In which direction are you heading? | direction | n,s,e,w |
| In what approximate direction is your destination? | direction | n,s,e,w |
| How far is it to your destination? | give # | keypad |
| How long will it take to get there? | say time | keypad |
| In what neighborhood/town/city are you now? | name it | x |

Table 9. IVSAWS questions.

| **Question** | **Voice Response** | **Button Press** |
|---|---|---|
| What should you do? | watch/pull over/ignore | 3 keys |
| Is it a warning for an emergency vehicle? | yes/no | y/n |
| Is it a slow moving vehicle? | yes/no | y/n |
| etc. | | |

Table 10. Vehicle monitoring questions.

| **Question** | **Voice Response** | **Button Press** |
|---|---|---|
| What should you do? | watch/pull over/repair soon/ignore/etc. | several keys |
| Could this damage the vehicle? | yes/no | y/n |
| Could this cause an accident? | yes/no | y/n |
| etc. | | |

The difficulty with these lists of questions is that not all information systems provide information to address all of these questions. For example, for the interfaces developed as a part of this research program, it was decided not to included destination distance or destination direction information, to simplify the interface.

A second problem with this method is that, in driving, information is accumulated over time. In response time experiments, it is presented in an instant. So, for example, an interface that simply said "right" or "left" (or showed arrows) when a turn was required would lead to very short response times. However, on the road, forewarning is required so the driver can be in the proper lane and driving at a speed at which the maneuver can be safely executed. While that problem could be overcome by using a series of slides to represent a trip, that makes implementing the test much more difficult. (See reference 50 for an example.) For some interface elements such as countdown bars and landmarks, dynamic testing may be required to accurately assess their utility.

Should such a test be conducted, the author recommends that it be conducted using a vehicle mockup, not just a screen and chair for the subject. Using the mockup helps put the subject in the proper frame of mind, though it is likely that, the ultimate effect on performance is small. It helps convince engineers who observe or read about the protocol, that the data are valid. If the engineers do not believe in the protocol, they are less likely to make the changes recommended by the human factors specialists, resulting in a driver interface that is not as safe or easy to use as it could be.

The hardware required for RT experiments consists of a personal computer, serial or analog interfaces, random access slide projectors, and response keyboards. Almost any personal computer will suffice. The cost for the computer and interface boards should be under $3500. Random-access slide projectors can cost about $2000 each.

The major cost for such systems is in integrating the hardware and developing the controlling software. Required software includes test code to exercise projectors and direct them to specific slide numbers, code to generate random and counterbalanced sequences of trials, code to schedule and process events in a trial (warning tones, projector movements, button presses, feedback to subjects), code to store responses, and code for the user interface to the entire program. The author spent about 6 months developing a highly flexible system for conducting such experiments. Once a general purpose response time program is written, modifying it for specific experiments is straightforward.

While this method is useful for evaluating the merits of particular features of visual displays, it is difficult to apply to overall assessments and to auditory or combined auditory/visual systems. For that reason, it is not recommended for system assessment.

## EXPERIMENTS USING VIDEOTAPE

Videotaped or filmed routes have often been used in driver training courses. Typically, drivers watch a tape for hazards, and when one appears, they step on the brake. In the ongoing project, videotaped segments were used to present the road scenes to drivers.[14] Concurrently, drivers were given route guidance information in various ways (visually or auditorily, and with and without landmarks) via a SuperCard simulation of the driver interface. The drivers' task was to step on the brake when the lead car on the videotape braked and press a button as soon as they could see the next maneuver referred to by the navigation display. This task was intended to cause drivers to pay attention to the road scene as they would while driving a real vehicle.

Conceptually, this method is appealing because it is inexpensive to set up and carry out, requiring only a videotaped journey, a VCR, and, of course, a navigation system synchronized to the tape. There were several problems with this approach. When the road scene was shown using a projection display people experienced motion sickness, likely accelerated because the in-car camera used to record the scene was not stabilized. Installing a gyro-stabilized camera was beyond the project budget. Presenting the road scene on a small monitor solved that problem but it changed the driver's accommodation distance, artificially shortening the time to look at the in-vehicle display and back at the road scene. However, the scene would still look unnatural and there are still some motion sickness problems induced by turns as drivers look in the direction of a turn before they make it. The unnaturalness could be overcome by aiming the camera to simulate driver head and eye movements (leading the turn), though the impact on motion sickness is uncertain. Also, the resolution of the video display was poorer than that of the real scene, which meant that drivers had to be closer to signs in the lab than on the road to read them. The eventual introduction of high-definition television (HDTV) may reduce this problem slightly, though its resolution falls far short of film and the resolving capability of the human visual system. Finally, the most significant problem was synchronizing the material on the tape with the computer controlling the IVHS display. This problem could be overcome by placing SMPTE time code on the videotape that can be read by the computer.

In terms of the experiment conducted as part of this project, only one lead vehicle was used, so the experiment could have become a car-following task. Multiple lead vehicles would solve this problem. In addition, there was a tendency for the driver of the lead vehicle to inadvertently cue the test subject to turns by moving to one side of the lane before a turn. Appropriate instructions to the drivers of lead vehicles will eliminate this behavior.

In spite of these weaknesses, the experiment was extremely convincing for participants. Drivers braked in response to the lead vehicle and some drivers turned the steering wheel when maneuvers were required (even though it had no effect on the videotape).

Recently, the author learned of plans for an improved version of this method, one which technology now permits.[61] Scenes from videotape are compressed in MPEG format and stored on a large hard disk (though they could be stored on a CD-ROM drive). When the driver reaches a decision point, any one of several routes can be displayed without interruption. Demonstration of this method may occur in the next year or so.

## DRIVING SIMULATOR TESTS

There are many who would advocate carrying out assessment tests in driving simulators. As was mentioned earlier, General Motors has three driving simulators, Ford has one, and Chrysler is developing a simulator. None of the suppliers have simulators nor do they have plans to develop them. In the academic arena, the University of Iowa has an excellent simulator and the National Advanced Driving Simulator (NADS) may be built there. Other simulators in the United States include those developed by Systems Technology Incorporated (at a few locations) and one being developed the contractor. Also, an alternative is the truck simulator sold by First Ann Arbor Corporation. The only Government simulator is at the Federal Highway Administration's Turner-Fairbank Highway Research Center in McLean, Virginia. It is particularly well suited for studies of highway signs.

The Department of Defense and defense contractors own many flight simulators which may not be needed in the future, as military spending decreases. Some could be converted to driving simulators, though at great cost. Many believe that because flight is three-dimensional and driving mostly two-dimensional, designing a flight simulator is more difficult. That is not true. The coupling of the car with the ground requires higher bandwidth for motion systems. Also, because the car is closer to the objects viewed (the ground), the rate of movement of key objects is greater.

Overseas, simulators are more prevalent. In Europe, BMW and especially Mercedes have good driving simulators, and many other car companies either have them or will have them shortly (e.g., Fiat). Many research organizations in Europe (VTI, TNO, University of Groningen, University of Stockholm, University of Leeds) either have or will shortly have excellent driving simulators. In Japan, several vehicle manufacturers have driving simulators (as was noted earlier) but suppliers developing IVHS driver interfaces do not have them.

At the present time, the author does not believe that final operational validity should be based on data collected in a driving simulator because it is not clear what needs to be simulated, how close the key features must be to reality, and how that depends upon the driver task. Second, there is little data in the literature indicating when simulator and on-road performance are correlated. Third, there are many conditions that most full-task driving simulators can not adequately represent, such as windshield distortion due to rain and the glare from oncoming headlights. Finally, cost considerations favor on-road testing.

It should be clearly understood, however, that simulators can play a fundamental role in preliminary tests. In determining if a simulator is adequate, the issue is "adequate for what." If the purpose of the simulation is to provide a loading task containing the attentional demands of driving, then a simulation that shows a curving road and traffic could be adequate. Simulators of this type might be used to assess the ease-of-use of car phones while dialing and engaging in conversation. Such simulators should also prove useful for assessing the usability of vehicle monitoring systems. If the purpose of the simulator is to assess IVSAWS interfaces, then low fidelity simulators may also be acceptable, though a wide field of view (with a rear scene) is preferred. If the purpose simulator is to provide the driver with a realistic traffic environment (to assess a traffic information system), then high traffic densities and four- and six-lane roads need to be shown. Very few of the simulators have that capability. If the purpose of the system is to examine use of a route guidance system then sign sight distance need to match those in the real world and landmarks need to be shown. Only the FHWA simulator can adequately represent signs, but presentation of landmarks needs improvement.

Simulators are particularly useful for comparing design concepts, especially during the early phases of a project. They provide a consistent test environment free from the uncertainties of weather conditions. They are well suited to exploring applications where on-the-road tests could impose significant risk to the driver (intelligent cruise control, collision avoidance, automated highways, etc.) However, the only true test of a driver information systems is on the road, and this is particularly true for navigation systems, where the necessary conditions cannot adequately be represented in a simulator.

# CONCLUSIONS

This report suggests that two procedures be used to evaluate the safety and ease-of-use of driver information systems. The main test procedure is to install either a real or simulated interface in an instrumented car, and drive over a specific, interim test route in southeastern Michigan. Data should be recorded to provide measures of the standard deviation of lane position, mean speed, speed variance, and the number and duration of eye fixations. In addition, performance measures for using the interface (such as the number of turn errors for a navigation system) should be obtained.

Once it is established that a basic system is safe and easy to use, it may not be necessary to retest the entire interface when only small changes are made. In that instance a simpler, less expensive, second procedure is suggested involving surveys of drivers at licensing offices. Those surveys will determine the percentage of the public that understands the interface. However, over a period of time the cumulative effect of changes may be substantial, and retesting on the road may be needed.

For all measures in both procedures three levels of acceptance are identified—best expected, desired/planned, worst case. On average, performance across measures should be at the desired/planned level and no performance measure should fall below the worst case level.

These suggested test procedures are offered along with the following caveats and assumptions:

- It is assumed that the Government will allow industry to carry out the bulk of the testing as it does with emissions and crash tests.
- It is assumed instrumented test vehicles can be readily developed.
- The data on which acceptance levels are set are weak, at best. Little is known as to what normal performance for the measures of interest is as a function of speed, lane width, traffic, and driver age.
- As an interim solution, a single site was chosen because the test roads are reasonably convenient to manufacturers, and there are some baseline data for driving them. Over the long term, alternatives need to be developed.

There are other potential test procedures such as response time studies, videotape protocols, driving simulator studies. However, for response time and videotape protocols, baseline data are lacking. For driving simulators, the fidelity requirements for an appropriate test are unknown. At the present time, high-fidelity driving simulators also may be cost prohibitive.

Missing from this report are methods to assess how driver interfaces affect the enjoyment of driving (referred to in various places as comfort/convenience/confidence). As was noted in reference 15, the preceding report, very little is known about how to measure those driver reactions, though physiological measures are a possibility. Driver ratings should also be considered. This is a topic in need of further study.

The author does not believe that the protocols and acceptance levels in their current form can be used for safety and usability certification procedures. However, this report should stimulate discussion and research on the specifics that could lead to protocols and acceptance levels that would prove to be useful for certification, and in their current form, should provide useful insights into assessment.

# REFERENCES

1 Green, P. (1993). *Human Factors of In-Vehicle Driver Information Systems: An Executive Summary* (Technical Report UMTRI-93-18), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

2 Green, P., Williams, M., Serafin, C., and Paelke, G. (1991). "Human Factors Research on Future Automotive Instrumentation: A Progress Report," *Proceedings of the 35th Annual Meeting of the Human Factors Society*, pp. 1120-1124.

3 Green, P. (1992). *American Human Factors Research on In-Vehicle Navigation Systems* (Technical Report UMTRI-92-47), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

4 Branda, J.E. (1990). *Attitudes Toward Advanced Automotive Display Systems: Feedback from Driver Focus Group Discussions* (Technical Report UMTRI-90-22), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

5 Green, P., and Brand, J. (1992). *Future In-Car Information Systems: Input from Focus Groups* (SAE Paper 920614), Warrendale, PA: Society of Automotive Engineers.

6 Green, P., Serafin, C., Williams, M., and Paelke, G. (1991). "What Functions and Features Should Be in Driver Information Systems of the Year 2000?" (SAE Paper 912792), *Vehicle Navigation and Information Systems Conference* (VNIS'91), Warrendale, PA: Society of Automotive Engineers, pp. 483-498.

7 Serafin, C., Williams, M., Paelke, G., and Green, P. (1991). *Functions and Features of Future Driver Information Systems* (Technical Report UMTRI-91-16), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

8 Green, P., and Williams, M. (1992). "Perspective in Orientation/Navigation Displays: A Human Factors Test," *Conference Record of Papers, the Third International Conference on Vehicle Navigation and Information Systems (VNIS'92)*, (IEEE Catalog #92CH3198-9), Piscataway, NJ: Institute of Electrical and Electronics Engineers, pp. 221-226.

9 Williams, M. and Green, P. (1992). *Development and Testing of Driver Interfaces for Navigation Displays* (Technical Report UMTRI-92-21), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

10 Paelke, G., Green, P., and Wen, C. (1993). *Development and Testing of a Traffic Information System Driver Interface* (Technical Report UMTRI-93-20), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

11 Serafin, C., Wen, C., Paelke, G., and Green, P. (1993). *Development and Human Factors Tests of Car Telephones*. (Technical Report 93-17), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

12 Hoekstra, E., Williams, M., and Green, P. (1993). *Development and Driver Understanding of Hazard Warning and Location Symbols for IVSAWS* (Technical Report UMTRI-93-16), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

13 Williams, M., Hoekstra, E., and Green, P. (1993). *Development and Evaluation of a Vehicle Monitor Driver Interface* (Technical Report UMTRI-93-22), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

14 Green, P., Hoekstra, E., Williams, M., Wen, C., and George, K. (1993). *Examination of a Videotape-Based Method to Evaluate the Usability of Route Guidance and Traffic Information Systems* (Technical Report UMTRI-93-31), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

15 Green, P., Hoekstra, E., Williams, M., George, K., and Wen, C. (1993). *Initial On-the-Road Tests of Driver Information System Interfaces: Route Guidance, Traffic Information, IVSAWS, and Vehicle Monitoring* (Technical Report UMTRI-93-32), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

16 Green, P., Hoekstra, E., and Williams, M. (1993). *On-the-Road Tests of Driver Interfaces: Examination of a Route-Guidance System and a Car Phone* (Technical Report UMTRI-93-35), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

17 Green, P. (1993). *Measures and Methods Used to Assess the Safety and Usability of Driver Information Systems* (Technical Report UMTRI-93-12), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

18 Green, P. (1993). *Suggested Procedures and Acceptance Limits for Assessing the Safety and Ease of Use of Driver Information Systems* (Technical Report UMTRI-93-13), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

19 Levison and Cramer. (1993) *Description of the Integrated Driver Model* (BBN Report 7840), Cambridge, MA: Bolt, Beranek, and Newman.

20 Levison, W.H. (1993) *A Simulation Model for the Driver's Use of In-Vehicle Information Systems* (TRB Paper 930935), paper presented at the Transportation Research Board Annual Meeting, Washington, DC., January 10-14, 1993.

21 Robertson, J., and Southall, D. (1992). *Summary Report on the Feasibility of Developing a Code of Practice for Driver Information Systems*, Loughborough, Leicestershire, United Kingdom: ICE Ergonomics Ltd.

22 Anastasi, A. (1968). *Psychological Testing* (3rd ed.), New York: MacMillan.

23 Zaidel, D.M. (1991). *Specification of a Methodology for Investigating the Human Factors of Advanced Driver Information System* (Technical Report TP 11199 (E)), Ottawa, Canada: Road Safety and Motor Vehicle Regulation, Transport Canada.

24 U.S. Department of Defense (1989). *Human Factors Engineering Design Criteria for Military Systems, Equipment and Facilities* (Military Standard MIL-STD-1472D), Philadelphia, PA: Naval Forms and Publications Center.

25 Pew, R.W. (1992). personal communication.

26 National Institute for Occupational Safety and Health (1981). *Work Practices Guide for Manual Lifting*, Cincinnati, OH: U.S. Department of Health and Human Service, Public Health Service, Centers for Disease Control, National Institute for Occupational Safety and Health, March.

27 Garg, A., Chaffin, D.B., and Herrin, G.D. (1978). "Predication of Metabolic Rates for Manual Materials Handling Jobs," *American Industrial Hygiene Association Journal*, 40, pp. 894-903.

28 Gould, J.D. and Lewis, C. (1985). Designing for Usability: Key Principles and What Designers Think, *Communications of the ACM*, March, 28(3), 300-311.

29 Good, Michael, Spine, T.M., Whiteside, J., and George, P. (1986). "User-Derived Impact Analysis as a Tool for Usability Engineering," *Human Factors in Computing Systems, CHI'86 Proceedings*, New York, NY: Association for Computing Machinery, pp. 241-246.

30 Callahan, J.M. (1986). "Ford Finds 400 Ways to Say Quality," *Automotive Industries*, January, 44-45.

31 Callahan, J.M. (1986). "Ford's Taurus Best-In-Class List: Part Two," *Automotive Industries*, February, 97-98.

32 Callahan, J.M. (1986). "Ford's Taurus Best-In-Class List: Part Three," *Automotive Industries*, September, 115-116.

33 McElroy, J. (1987). "For Whom Are We Building Cars?" *Automotive Industries*, June, 68-70.

34 Hauser, J.R., and Clausing, D. (1988). "The House of Quality," *Harvard Business Review*, May-June, 63-73.

35 Garber, N.J., and Gadiraju, R. (1989). "Factors Affecting Speed Variance and Its Influence on Accidents," *Transportation Research Record #1213*, pp. 64-71.

36 Sweet, R.E., and Green, P. (1993). UMTRI's Instrumented Car, *UMTRI Research Review*, January-February, 23(4), 1-12.

37 Green, P., Boreczky, J., and Kim, S. (1990). "Applications of Rapid Prototyping to Control and Display Design." (SAE Paper #900470, Special Publication SP-809), Warrendale, PA: Society of Automotive Engineers.

38 Godthelp, H., Milgram, P., and Blaauw, G.J. (1984). "The Development of a Time-Related Measure to Describe Driving Strategy," *Human Factors*, June, 26,(3), 257-268.

39 Godthelp, H. (1988). "The Limits of Path Error-Neglecting in Straight Line Driving," *Ergonomics*, 31(4), 609-619.

40 Verwey, W.B. (1991). *Towards Guidelines for In-Car Information Management: Driver Workload in Specific Driving Situations* (Technical Report IZF 1991 C-13), Soesterberg, The Netherlands: TNO Institute for Perception.

41 Noy, Y.I. (1990). *Attention and Performance While Driving with Auxiliary In-Vehicle Displays* (Transport Canada Publication TP 10727 (E)), Ottawa, Ontario, Canada: Transport Canada, Traffic Safety Standards and Research, Ergonomics Division.

42 Levison, W. (1993), (personal communication).

43 Godthelp, H. and Kappler, W.D. (1988). "Effects of Vehicle Handling-Characteristics on Driving Strategy," *Human Factors*, 30, 219-229.

44 Carson, J.M. and Wierwille, W.W. (1978). "Development of a Strategy Model of the Driver in Lane-Keeping," *Vehicle Systems Dynamics*, 7, 223,253.

45 Noy, Y.I. (1990). *Attention and Performance While Driving with Auxiliary In-Vehicle Displays* (Transport Canada Publication TP 10727 (E)), Ottawa, Ontario, Canada: Transport Canada, Traffic Safety Standards and Research, Ergonomics Division.

46 Walker, J., Alicandri, E., Sedney, C., and Roberts, K. (1991). "In-Vehicle Navigational Devices: Effects on Safety and Driver Performance," *Public Roads*, 56(1), 9-22.

47 Green, P., Lin, B., and Bagian, T. (1993). *Effect of Road Geometry on Driving Workload: A Pilot Experiment* (Technical Report UMTRI-93-39), Ann Arbor, MI: The University of Michigan Transportation Research Institute (in preparation).

48 MacAdam, C. (1993), (personal communication)

49 Allen, R.W., and McRuer, D.T. (1977). *The Effect of Adverse Visibility on Driver Steering Performance in an Automobile Simulator* (SAE Paper 770239), Warrendale, PA: Society of Automotive Engineers.

50 Jex, H.R., Allen, R.W., DiMarco, R.J., and McRuer, D.T. (1974). *Alcohol Impairment of Performance on Steering and Discrete Tasks in a Driving Simulator, Part I: Effects of Task Loading, Part II: Moderate vs. Heavy Drinkers* (Technical Report DOT-HS-801-302), Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration.

51 Godthelp, H. (1988). "The Limits of Path Error-Neglecting in Straight Line Driving," *Ergonomics*, 31(4), 609-619.

52 Zwahlen, H.T. and DeBald, D.P. (1986). "Safety Aspects of Sophisticated In-Vehicle Information Displays and Controls," *Proceedings of the Human Factors Society-30th Annual Meeting-1986*, Santa Monica, CA: The Human Factors Society, pp. 256-260.

53 Wierwille, W.W., Antin, J.F., Dingus, T.A., and Hulse, M.C. (1988). "Visual Attentional Demand of an In-Car Navigation Display," Gale, A.G. et al (eds.), *Vision in Vehicles II*, Amsterdam, The Netherlands: Elsevier Science, pp. 307-316.

54 Senders, J.W., Kristofferson, A.B., Levison, W.H., Dietrich, C.W., and Ward, J.L. (1967). "The Attentional Demand of Automobile Driving," *Highway Research Record #195*, Washington, DC: Highway Research Board, 15-32.

55 Streeter, L.A., Vitello, D., and Wonsiewicz, S.A. (1985). "How to Tell People Where To Go: Comparing Navigational Aids," *International Journal of Man-Machine Studies*, 22, 549-562.

56 Schraagen, J.M.C. (1990). "Use of Different Types of Map Information for Route Following in Unfamiliar Cities," van Winsum, W., Alm, H., Schraagen, J.M. and Rothengarter, T., *Laboratory and Field Studies on Route Representation and Drivers' Cognitive Models of Routes* (Deliverable GIDS/NAV2), Haren, The Netherlands: University of Groningen, Traffic Research Center.

57 Antin, J.F., Dingus, T.A., Hulse, M.C., and Wierwille, W.W. (1990). "An Evaluation of the Effectiveness and Efficiency of an Automobile Moving-Map Navigational Display," *International Journal of Man-Machine Studies*, November, 33(5), 581-594.

58 Green, P. and Miller, D. (1983). *A Human Factors Evaluation of a Vehicle Maintenance Monitor.* Videotape report produced for Ford Motor Company (Electrical & Electronics Division and Automotive Safety Office), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

59 Paelke, G., and Green, P. (1993). *Development and Testing of a Traffic Information System Driver Interface* (Technical Report UMTRI-93-20), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

60 Saunby, C.S., Farber, E.I., and DeMello, J. (1988). *Driver Understanding and Recognition of Automotive ISO Symbols* (SAE Paper 880056), Warrendale, PA: Society of Automotive Engineers.

61 Kantowitz, B. (1993), personal communication.