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

NHTSA
www.nhtsa.gov

DOT HS 811 763 May 2013

Support for NHTSA Visual-Manual Guidelines: Expert Review of the Visual Occlusion Method and How It Compares to Driver Eye-Glance Behavior

DISCLAIMER

REPORT DOCUMENTATION PAGE		*Form Approved* *OMB No. 0704-0188*
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE May 2013	3. REPORT TYPE AND DATES COVERED
4. TITLE AND SUBTITLE Support for NHTSA Visual-Manual Guidelines: Expert Review of the Visual Occlusion Method and How It Compares to Driver Eye-Glance Behavior		5. FUNDING NUMBERS
6. AUTHOR(S) Miguel Perez, Melissa Hulse, and Linda Angell		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Virginia Tech Transportation Institute 3500 Transportation Research Plaza (0536) Blacksburg, VA 24061		8. PERFORMING ORGANIZATION REPORT NUMBER DOT HS 811 763
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Department of Transportation National Highway Traffic Safety Administration 1200 New Jersey Avenue SE. Washington, DC 20590		10. SPONSORING/MONITORING AGENCY REPORT NUMBER
11. SUPPLEMENTARY NOTES		
12a. DISTRIBUTION/AVAILABILITY STATEMENT Document is available to the public from the National Technical Information Service www.ntis.gov		12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The occlusion method is one of the verification test alternatives allowed under the Alliance of Automobile Manufacturers guidelines to manage visual demand of certain in-vehicle devices. Since the time those guidelines were published, additional research has become available on the advantages, disadvantages, and considerations of using the occlusion method in this context. The current investigation reviewed this additional research and then asked several occlusion experts about their opinions regarding topics such as method parameter values, test settings, validity, and usefulness of results. Experts' answers were summarized and suggested that there is consensus about some aspects of the occlusion method and disagreement over other aspects. These findings are discussed in the context of including the occlusion method as a test alternative in future versions of visual-manual device design guidelines.

14. SUBJECT TERMS			15. NUMBER OF PAGES
Driver distraction, evaluation methods, visual demand, secondary task, in-vehicle tasks, design standards, guidelines, occlusion, in-vehicle information systems			56
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

NSN 7540-01-280-5500 Standard Form 298 (rev. 2-89)
Prescribed by ANSI Std. 239-18, 298-102

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND SYMBOLS

IRB	Internal review board
ISO	International Standards Organization
IVIS	In-vehicle information system
JAMA	Japanese Automobile Manufacturers Association
R metric	Resumability ratio
RL	Resumption lag
SCT	Shutter closed time
TSOT	Total shutter open time
TGT	Total Glance Time
TTT	Total task time
VCR	Video cassette recorder

CHAPTER 1. INTRODUCTION

BACKGROUND

The National Highway Traffic Safety Administration (NHTSA), as part of its "Distraction Plan," is in the process of developing guidelines for devices that may be sources of distraction for drivers. As part of this effort, several knowledge gaps have been identified. One of these knowledge gaps, which has implications for the development of these guidelines, is whether the occlusion method should be included as a test alternative. A set of guidelines issued by the Alliance of Automobile Manufacturers (Driver Focus-Telematics Working Group, 2006) allows for the use of the occlusion method, but some research has emerged that points to potential issues with the method.

To examine these issues, a small sample of experts were interviewed. These experts were intimately familiar with the occlusion method. Interview participants were prompted to provide input that resulted in the recommendations that are presented within this document.

SCOPE

The first step in this project was to examine the recent relevant research on the occlusion method. In particular, research comparing occlusion and measures of visual demand was of interest for this review. Following the literature review, questions were generated to prompt experts for their opinions about whether the occlusion metric should be included in the NHTSA Visual-Manual guideline and, if recommended for inclusion, whether adjustments to the metric or its parameters should be considered. Experts were also asked to provide feedback on unique advantages and disadvantages of the occlusion method and were free to contrast occlusion with traditional eye-glance observation.

CHAPTER 2. RECENT LITERATURE REVIEW FOR THE OCCLUSION METHOD

The occlusion method was originally developed by John Senders in the 1960s as a way to better understand the visual demand of primary tasks associated with driving a vehicle (see, for example, Senders, Kristofferson, Levison, Dietrich, & Ward [1966, 1967a, b] and Senders & Ward [1968]). In the last decade, however, the method has been applied to secondary device use, yielding a measure of the visual demand for in-vehicle devices. Visual demand is an important measure for in-vehicle devices given that higher levels of measured visual demand (however it may be measured) imply a higher crash risk if a task is performed concurrently with driving. The occlusion method is one of those methods that has been proposed as a surrogate for more expensive and complex simulator, test-track, and on-road tests, which require that drivers' eye glances are recorded and later analyzed by human observers, or assessed using eye-tracking technologies. It is important to note that the occlusion method can only be applied to visual-manual tasks. It is not valid for use with auditory-vocal-cognitive tasks, nor with mixed-mode tasks.

Various procedural specifications have been developed for the occlusion method. Those include a standard from the International Standards Organization (ISO 16673, "*Occlusion method to assess visual demand due to the use of in-vehicle systems,*" 2007) and guidelines from the Alliance of Automobile Manufacturers (Driver Focus-Telematics Working Group, most recently updated in 2006), the Japanese Automobile Manufacturers Association (JAMA, 2003), and SAE J2364 200408 ("*Navigation and route guidance function accessibility while driving*," Society of Automotive Engineers, 2004). These standards and guidelines were mostly based on research conducted before 2005. Foley (2008) summarizes the standards well:

> "ISO standard 16673 provides consensus on the visual occlusion procedures. The specific criterion value (to aid in the decision of whether or not the task will be available to the driver while driving) is established by the user of the standard. Some criteria have been published. The Alliance guidelines state that [Total Shutter Open Time] TSOT, for a task accessible to the driver while driving, should not exceed 15 sec. The JAMA guidelines (Section 4.2) state that 'The visual information to be displayed shall be sufficiently small in volume to enable the driver to comprehend it in a short time or shall be presented in portions for the driver to scan them in two or more steps' (p.2). The occlusion method is included in the verification techniques in Annex 3 to the JAMA guidelines, stating that the total shutter opening time (with vision intervals of 1.5 sec and occluded internals of 1.0s) should be no more than 7.5 sec. This criterion is based on the observed correlation between TSOT and total glance time (TGT), measured using direct eye glance performance (Karlesson & Fichtenberg, 2001; Hashimoto & Atsumi, 2001, p.130)."

These standards were based on the research available prior to their development and publication. Additional research has become available since the standards were published, and is highlighted in this summary. The research is organized chronologically. For ease of reading, the initial mention of each study is made in bold-faced type.

Baumann, Keinath, Krems and Bengler (2004) concluded, based on their empirical data, that the occlusion technique was a reliable and valid method for evaluating visual as well as dialogue aspects for in-vehicle information systems. In general, their results showed that the occlusion approach was, in terms of total task time and number of errors, more representative of secondary task performance under actual driving than static task performance would be (in which undivided attention was devoted to the secondary task). Baumann et al. (2004) collected data on 43 drivers who performed in-vehicle tasks using an on-board information system. Thirteen of these participants drove the vehicle on actual roads while inputting 15 destinations. Sixteen drivers did not drive, but sat in the driver's seat and inputted the same 15 destinations while the vehicle was static. The remaining 14 drivers also sat in the driver's seat, but entered the 15 destinations while wearing occlusion goggles. The goggles cycled between occlusion periods lasting 1.5 seconds and shutter-open periods lasting 3 seconds. These times were considered to represent the maximum time tolerance range for eyes-off-road time taken from research done by Zwahlen et al. (1988). The destinations entered had three different input lengths: short (5 to 8 inputs), medium (9 to 11 inputs), and long (12 to 13 inputs). The dependent variables were total task time (TTT) and the number of errors.

Their results showed that the static condition had the fastest TTT and fewest errors. Drivers who had to drive and perform the tasks took considerably longer to complete the task and had the most errors. The occlusion condition showed intermediate performance. Occluded drivers took less time than the on-road drivers but significantly more time than the static (but unoccluded) drivers. The same was true for the number of errors.

When comparing TTT and calculating a scaling factor of how much longer one input condition was compared to another, they found a multiplicative ratio of 1.82 between static and occluded task completion times for short inputs, a multiplicative ratio of 2.19 between driving and static task completion times for medium inputs and a multiplicative ratio of 1.27 between the driving and occlusion task completion times for long inputs. For the error data, the differences were 1.31 times more errors between static and occluded conditions for short inputs, 1.29 between driving and occluded conditions for medium inputs, and 1.7 between driving and static conditions for long inputs. However, the authors stress that these scaling factors were specific to their study and cannot be easily transferred to other settings.

Noy, Lemoine, Klachan, and Burns (2004) examined task interruptability using the occlusion technique Task interruptability and resumability, R, was proposed to measure how easily the information processed from a prior glance is remembered when the task is resumed. R is calculated as the mean TSOT divided by TTT, the latter calculated in a trial where the participant's vision remains unoccluded throughout the task. A value for R of 1 (or less) is considered by proponents of the R measure to indicate that the task is very interruptible and resumable without undue costs or penalties to performance.

Twenty-four volunteers participated in the study. Participants first performed a radio-tuning task and two visual search tasks when their vision was not occluded, followed by the same type of tasks with intermittent occlusion (1.5 second viewing time, 3.0 second occluded time). This was followed by a second session involving driving in a simulator and performing similar tasks again. While the total task time while unoccluded was not sensitive to different tasks, there were significant differences between tasks for TSOT and R in the occluded condition. A scrolling task, which was impossible to perform during an occluded interval, had high TSOT and R values. In turn, the radio-tuning task had correspondingly lower TSOT and R values, since it could continue to be performed during the occluded interval. Noy et al. (2004) reported that these findings were similar to those of Krems, Keinath, Baumann, Bengler, and Gelau (2000) where an unpredictable task, such as the scrolling screen, caused more errors than a predictable task. The tasks in Noy et al. (2004) were found to be sufficiently difficult to affect driving performance in a simulator. In addition, there were some noteworthy correlation levels between the occlusion measures and lane-keeping performance. Lane-keeping measures were substantially correlated with TSOT for the harder visual search tasks. Likewise, the data suggested a potential correlation between lane keeping measures and the duration of a task. Finally, drivers reported that mental and temporal demand, effort, frustration and safety were more difficult when tasks were performed while driving than during the occlusion trials.

Based on these results, Noy et al. (2004) could not suggest the use of one metric (TSOT or R) over the other. Tradeoffs between the measures exist. The authors do suggest, however, that both R and TSOT are needed:

> "Perhaps even a composite measure such as $TSOT^2/TTT_{unocc}$." (p. 212)

Asoh and Iihoshi (2006) presented data from various experiments designed to understand the relationship between Total Glance Time (TGT) and other task assessment tools, including the occlusion method. They first measured TGT, lane deviation, and participant uneasiness for performing different navigation-related tasks. They compared TGT while performing tasks for different roadway types (2-lane urban, 1-lane urban, Joban Expressway, Metropolitan Expressway) and IVIS input devices (Touch Panel, Joy Stick, Remote Control, Rotate). The authors found the lowest TGT (7.9 seconds) to correspond with conditions in which 1) participants had neutral feelings concerning uneasiness of performing the task, and 2) lateral control was not affected.

Further findings reported by Asoh and Iihoshi (2006) suggested that TSOT had a stronger correlation with TGT (0.8933) than TTT (0.8760) for bench testing. The bench testing intervals for closed and open occlusion times were:

- Closed 0.5 seconds, open 1.0 seconds
- Closed 1.0 seconds, open 1.0 seconds
- Closed 0.5 seconds, open 1.5 seconds
- Closed 1.0 seconds, open 1.5 seconds
- Closed 3.0 seconds, open 1.5 seconds

The interval exhibiting the highest correlation was when the shutters were closed 1.0 seconds and open 1.5 seconds. Based on these data, a TSOT of 7.5 seconds was found to be equivalent to 8 seconds of TGT, which in turn was the TGT that resulted in neutral uneasiness levels.

Angell, Auflick, Austria, Kochhar, Tijerina, Biever, Diptiman, Hogsett, and Kiger (2006) conducted a review of various rules proposed for whether tasks were safe to complete while driving (as part of a larger program of work on Driver Workload Metrics). They first established a "ground truth" classification of whether the tasks to be tested were lower or higher workload tasks (by combining three sources of information: (i) findings from the prior scientific literature about each task's workload, (ii) predictive modeling of each task, and (iii) engineering judgments by experts in the field of driving and distraction about the workload level of each task). They then ran a study having participants perform those tasks (in each of three venues: a static laboratory environment, a test track, and on the road), in which they obtained a number of measurements. In the laboratory, these assessments included the use of visual occlusion, and the measurement of time needed to complete tasks when vision was occluded and also when no occlusion goggles were used. Angell et al. (2006) compared the findings to the a-priori "ground truth" workload levels established for each task. The authors found that the best correspondence between different task classification rules, based on various measures of visual demand and the a-priori "ground truth" workload levels, occurred when a TSOT threshold of 7.5 seconds was used. Angell et al. (2006) also showed correlations higher than 0.8 between TSOT and TGT. These findings corresponded well with those of Asoh and Iihoshi (2006), described above. Median TSOT was found to correlate significantly with both track and while driving on a road driving performance measures: median task duration, median standard deviation of lane position, lane exceedances, median speed difference, number of glances to task, and total glance time (TGT) to task. Median TSOT did not correlate with glance duration, or with event detection measures. However, the R metric from occlusion did not discriminate well between the visual demand of high and low workload tasks, and it also did not correlate at a significant level with any driving performance measures (in other words, it had no predictive validity and was not able to discriminate between levels of visual demand). However, the authors found that the visual occlusion metric of TSOT did lead to repeatable results (split-half correlation, r = 0.996), with predictive validity for visual demand on the road and track (based on the many significant correlations reported above), and discriminated effectively between high and low workload tasks for visual-manual tasks (88% correct discriminations).

Monk and Kidd (2007) had participants complete Video Cassette Recorder (VCR) tasks using the occlusion method while performing a tracking task (following an airplane using a mouse cursor). The results showed that the R value was smaller for VCR tasks that included a tracking task. The authors conclude that the R value may be overestimated when participants do not have a tracking task to complete during the occlusion interval. It is also reported that:

> "Estimations of visual demand based on individual post-occlusion resumption times may provide a more precise measure of transition costs and resumability than measures based on Total Shutter Open Time." (p. 2).

As a result, Monk and Kidd (2007) propose the following formula for evaluating visual demand costs:

$$D = \frac{(\overline{RL} - \overline{IAI})\, n}{TSOT}$$

where RL are the resumption lags and IAI are the times between unoccluded task inputs during each unoccluded episode. For their study, R and D measures were not highly correlated. The authors state that this might be due to the small sample size of eight participants. They suggest that D may provide a more precise estimate of resumability because it is based on resumption measures.

Horberry, Stevens, Burnett, Cotter, and Robbins (2008) stated that guidelines should be developed for a detailed measurement protocol using standard occlusion. They specifically addressed the issue of age in relation to In-Vehicle Information System (IVIS) tasks. The ISO standard states that 20 percent of participants should be over the age of 50 when testing using the occlusion method. The authors believe that more research is needed in this area. In their study, 60 participants ranging in age from 17 to 76 completed 4 different tasks ranging from easier (entering a point of interest – approximately 4 key presses) to harder (long scrolling – finding a 3 letter stock code from a scrolling list and reading out the accompanying share price).
The results showed differences in age with respect to TSOT and R. The older participants had more varied scores, especially for TSOT, and generally did more poorly than younger participants. The authors therefore suggest that to obtain minimal inter-subject variability measurements, younger/middle-aged participants (broadly 17-66 years old) should be used. The authors emphasize that this range merely reduces variability during testing and that the recommendation should not be taken as a suggestion that IVIS systems are developed only for younger drivers. Note that this age range differs from the one recommended for use in testing by the Alliance (45-65 years of age).

Gelau, Henning, and Krems (2009) looked at how reliable the occlusion technique is for assessing visual distraction with IVIS. They looked at 4 different experiments that were previously conducted and examined the internal consistency of the TSOT measure. Their findings showed that the:

> "Characteristic values for internal consistency were almost all in the range of 0.90 for the occlusion method and can be considered satisfactory." (p.184)

Brook-Carter, Stevens, Reed, and Thompson (2009) looked at issues related to research gaps reported by Stevens et al. (2004). They were:

- System response delay
- Participant training levels prior to a study
- Number of repetitions for each task
- Sample size required
- Calculating resumability

The study included 20 participants ranging in age from 28 to 55. The four tasks each participant completed were destination entry, place of interest search, changing radio frequency, and dialing a telephone number. All participants did five training sessions on each task followed by three testing trials for each task. Half of the participants used occlusion goggles during the training session and the other half did not. All participants used the occlusion goggles during the testing trials.

The authors reported that system response delays were present in the systems and tasks used in the experiment and that they could vary a great deal between different attempts at both the same and different tasks[1]. Based on these findings, the authors suggest that three or more training

[1] The ISO Standard 16773 on visual occlusion contains a special provision for the treatment of system response delays.

tasks should be used for the more complex tasks being tested (e.g., place of interest search). Tasks that are less complex would require fewer training sessions. The results also showed no significant difference with respect to the occluded versus non-occluded training method. Therefore, training with or without occlusion goggles shouldn't have a difference in the outcome of the research findings. In addition, based on differences between task completion times for successive trials of the same task, the authors recommend the use of two trials. In terms of sample size, the results suggested the use of over 10 participants for some tasks, although other tasks would benefit from a minimum sample size of 14 participants. The authors suggest a potential tradeoff of increasing the number of test trials rather than the sample size, which could be less time-consuming and expensive.

Related to statistics, Brook-Carter, et al. (2009) also suggest that:

> "The value of R should be calculated on a within-participant basis by taking the ratio of the respective median TSOT and TTT_{static} across trial repetitions completed by one participant on each task under evaluation. Also, for the purpose of identifying interface designs that exhibit poor resumability, the 85^{th} percentile value is suggested as the most appropriate (p. 185)."

Gelau and Schindhelm (2010) conducted a study having participants perform real driving tasks along with a simple auditory tracking task. Their findings supported those of Monk and Kidd (2007), discussed earlier. There was a significant increase in the TSOT and R values when participants performed the auditory tracking task. TSOT and R value measures obtained when there was a concurrent auditory tracking task also showed a clearer discrimination between an "easy" versus a "difficult" IVIS task than measures obtained when there was no concurrent auditory tracking task. Furthermore, the R value was found to only be sensitive to the level of difficulty for the IVIS task when the auditory tracking task was used. The authors propose amending the standard occlusion method by having it include a cognitive load task, which they believe would improve the sensitivity of the method.

Finally, **Stevens, Burnett, and Horberry (2010)** point out that the current metric for the occlusion method lacks validation and does not:

> "Constrain measurement variability as much as it could" (p. 530).

The authors believe the ISO standard is a good starting point but does not include sufficient measurement detail. They proposed a new metric to test IVIS tasks that does not use R:

$$\text{Mean TSOT} + (85^{th}\text{ percentile TSOT} - \text{Mean TSOT})^2/\text{Mean TSOT} < 8.0 \text{ seconds}$$

They justify the use of the 85^{th} percentile because occlusion data are not usually normally distributed but instead follow a log-normal curve (Green, 2002; this observation is also justified by the Stevens et al., 2010, study). It is noteworthy that the authors used the data from Horberry et al. (2008), whose drivers ages ranged from 17 to 76 years of age. Other studies reported in this literature review mainly used younger-aged participants. Stevens et al. (2010) state that using the 85th percentile allows for more representation from participants who had more difficulty with some tasks. The selection of TSOT is based on the most conservative measure available, which is JAMA's mean value of 7.5 seconds. The authors adjusted that number to 8.0 seconds to account for both the mean and the spread of the data.

A tabular summary of the literature review presented in this section was provided to the experts for review prior to the interview. This was intended to allow the experts to familiarize themselves with recent findings related to the occlusion method and indirectly capture their reactions to the criticisms of the occlusion method parameters and approach that some of these studies express. It should be noted that some additional research on visual occlusion – beyond that reviewed here – has been done, but not published (and thus is not readily accessible). Some of it has been reported *informally* at meetings of ISO Working Group 8, and/or at the various Driver Metrics Workshops that have been held (often co-sponsored by SAE and the Alliance). However, many of the experts who were interviewed for this project are participants in these groups – and thus have had some exposure to emerging work beyond that which has been published. The interview process itself, and the results obtained, are described in the next chapter.

CHAPTER 3. EXPERT INTERVIEWS

INTERVIEW PROTOCOL

Potential participants were sent an e-mail invitation and RSVP form (with informed consent statement included) which they returned with a response to the invitation (see Appendix A). Follow-up phone calls were made in some cases. Those experts that indicated they were interested in participating in a phone interview were contacted/scheduled for a phone interview. Before the scheduled date of the interview, a packet was emailed to interviewees, containing a table with a summary of recent literature on the occlusion method (see Appendix B). This packet was kept by the interviewees. Interviews were either recorded or notes were written concerning the interviewees responses. The interviewees decided whether they wanted their answers recorded or notes to be taken. The interviews lasted approximately one hour. Interview scripts are included in Appendix C.

The interviewer read from the scripts, and provided clarification as necessary during the interviews. Also, the interviewer occasionally asked additional follow-on questions as they naturally arose during the conversation. The script was used as a guideline and any additional clarification or follow on questions were consistent with the subject matter and research questions of interest.

For those participants who chose to complete a written questionaire, the written questionnaire (Appendix D) was sent via e-mail after their signed inform consent was received. The questionnaire was expected to take approximately 1 to 2 hours to complete. Questionnaires were returned to the researchers via e-mail.

PARTICIPANTS

Eleven interview requests were originally sent. From these requests, six experts agreed to be interviewed for this study. Three experts declined the opportunity to participate, mainly because they were no longer performing work on occlusion. One expert agreed to complete the written interview but did not respond once it was sent, despite multiple reminders. One expert never responded to the original invitation. From the six experts who participated, three were employed by automotive manufacturers and three were employed by research or academic institutions. Four of the experts were international, three of these from Europe.

SUMMARY OF EXPERT RESPONSES

The experts provided substantial insight on all the questions that they were asked to answer. Their answers are summarized in this section, including areas of concordance and disagreement in their opinions.

Are you familiar with the Alliance treatment of the occlusion method?

All six participants responded that they were familiar with the Alliance treatment of the occlusion method.

Do you think the occlusion metric of Total Shutter Open Time (TSOT) is sufficient to describe visual demand for verification purposes? Why or why not?

Of the six respondents, only one replied no. He replied:

> "No. From my point of view, TSOT as defined by ISO 16673 (p.3) reflects the costs of task interruptions if the task consists of subtasks requiring single glances of more than 1.5, cannot be blindly operated, and/or task resumption is complicated. By the R metrics these costs are expressed independently from Total Task Time. Interruptability and chunkability are important aspects at the assessment of visual demands. Nevertheless, for a full description of visual demands the occlusion metrics should be supplemented by direct measure."

Another respondent replied that the occlusion technique is a sufficiently valid method to measure visual demand:

> "First of all: Define your purpose: If your purpose is to have a valid method and an existing measurement protocol that is easy to use, the occlusion method will lead to good results."

He also stated that:

> "If you only accept field experiments or driving simulator studies with eye tracking as valid you will tend not to accept the occlusion technique. However you will be stuck in that case with a much more complicated methodology and procedure."

Following along this theme was the following response:

> "The method was developed to be independent of technology; not dependent on where technology is located, or which technology is used. Considering traffic demands, visual behavior, this is coming from classical studies based on empirical data, still a valid measurement of visual demand."

Two respondents stated that the occlusion method can help verify the visual demand of driver-vehicle interfaces, but it should not be the only assessment tool. One expert said:

> "It should be one tool in a battery of assessment methods. It should be part of a staged assessment protocol, or triage type approach, to limiting tasks that are not compatible with safe driving."

The second expert indicated:

> "Distraction is too complex for any single metric to complete[ly] describe what is happening to the driver."

One respondent commented on the R metric, stating:

> "I do not understand or believe that the R metric is appropriate. This is because we looked into the R metric in the past and we have not found any correlation from it to on road driving and do not really understand what it's measuring to be able to properly explain it."

These responses suggest that the experts were comfortable with using the occlusion method as a tool to assess visual demand, but preferably in the company of other methods. As one expert said:

> "The main limitation of occlusion is that it does not directly address single glance time. It does demonstrate that the task can be completed without excessively long glance times, but not what a driver would actually do if the task was attempted while driving."

Should the parameters for the occlusion test be adjusted? Why?

Three respondents felt that the current parameters were fine as is. One of those felt that:

> "If there is sufficient new data available to rewrite the document, start a new ISO process for revision of the document. Do not adjust the parameters in another guidelines document, try to establish one document (ISO) and refer to this."

Two experts felt that participants should perform a task during the shutter closed period. One stated that:

> "The evidence from Gelau & Schindhelm (2010), that task load during the shutter closed provides a more sensitive method for discriminating meaningful differences in tasks, suggests that we should load the shutter closed interval."

It was also suggested that the:

> "1.0 second shutter closed interval is too short. I was not keen on the 1.5 second period that ISO selected. A 2.0 second shutter closed interval would be more effective and realistic. I do not have the evidence at hand to reference in support of this. If the glance behavior is to be even remotely comparable to real driving, the eyes-on-road time (shutter closed) should exceed eyes-off-road time (shutter open). This is not the case with the 1.5 seconds open and 1.0 second closed intervals."

Another expert also stated that:

> "Total task time is highly correlated to total glance time – for visual manual tasks. There is a lot of information in these relationships. I think the criterion with a limit of 15 seconds is pretty short for comparing it to tasks that are successfully performed in cars. The question is whether one should add the interruptability criterion."

In reference to the R metric, one expert stated:

> "I would argue that the R metric should not be used. TSOT is not a perfect measure of task distraction risk but it does capture attributes of visual demand. Problems with R: it is not clear what R measures, R does not discriminate task demands/complexity well, increases trial demands, [and has] face validity issues."

This was also stated by a different respondent when answering the first question.

Finally, one respondent replied with:

> "The parameters of the occlusion procedure have been demonstrated to be somewhat elastic. Within reason one would get the same answer with different specific values. I think the most critical value is the shutter closed time (SCT). SCT has to be long enough (at least 1.0 second) to interrupt the ongoing task but not so long (> 2.5 seconds) so the subject becomes frustrated and unwilling to complete the procedure. My personal preference is for a SCT of 1.5 seconds to ensure more complete interruption of the task as with only 1.0 second there is less likelihood that short term memory will be exceeded."

In summary, there was limited agreement by experts on this particular parameter. Two experts used this question to express their belief that a secondary task should be added during the shutter closed interval. Apart from the increased perceived realism (as when one is driving and not conducting an in-vehicle task, the person is arguably attending to the task of driving), there seems to be some evidence that this manipulation can alter test results. More in line with the original question, there were comments on adjusting the shutter closed interval. Some experts felt the closed time interval should exceed the TSOT, another thought 1.5 second was sufficient and a third felt a 2.0 second SCT would be more effective and realistic.

Could there be an influence from limiting the maximum glance duration at 1.5 seconds?

One expert said no. A second said that:

> "The 2 seconds criteria must not be adjusted [Author's note: the expert is referring to the 1.5 shutter open time plus the assumed 0.5 transition time from the forward view, see below]. Stay with 2 seconds glance maximum. More experiments are needed to see if you can expand the 15 seconds criteria. Never expand the 2 second criterion. More and more new automotive content seems to be going to the limit of readability. The sensitive parameter to test this is the 2 second criteria. Length of interval – open interval – of 1.5 seconds is very discriminative to check the 2 second rule. Stay with 1.5 open. It corresponds to 2 seconds minus half a second from road to display at the beginning (before the glance begins). I wouldn't weaken this open time."

A key insight from this expert is that the 1.5 seconds shutter open interval was intended to correspond to a 2 seconds maximum duration glance minus the time it would typically take for a saccade at the beginning from the road to the device (which was estimated to be about 0.5 seconds).

Another expert stated that:

> "It might force designers to limit the amount of visual information/choices and steps in their tasks, find alternative modes of interaction (i.e., voice) or lock out tasks that can't meet the requirements. The 1.5 second limit may limit the diagnosticity of the procedure – the actual duration and distribution of glances needed to perform a task while driving will not be known and problem areas will not be specifically identified. You'll only know that there is an issue with visual demand. Again, this emphasizes the need to assess visual demands in several ways and not just with occlusion."

A different respondent said:

> "Yes, of course, but one of the reasons for that value is to provide evidence that a longer glance, which is potentially more distracting, is not needed to complete the task."

One expert replied that he has looked at different open periods, including extremely short ones, and for different tasks, with the conclusion that 1.5 seconds is reasonable.

Overall, it seems as though the experts felt that 1.5 seconds was reasonable. It was pointed out that limiting this value further may influence the diagnosticity of the procedure and that visual demands need to be assessed in several ways to get a full picture of what a driver is capable of contending with. At the very least, there was no indication that the experts believed that the interval should be shortened.

Could there be an influence from adjusting the occluded internal from 1.0 second to 1.5 seconds?

One expert indicated that:

> "The longer you make that period; I think people just get a little more anxious with it because it takes a bit more time to perform tasks. However, it does also give folks the ability, once they have become comfortable with the occlusion procedure, to have more time to perform actions and inputs while the goggles are closed. So over a period of time, you may see that participants with a longer occluded period could perform tasks with a – with fewer open periods. But again, I think there's some research out previously that showed there was not a difference, but if they are properly trained on it, there may – you may find a difference. So increasing the occluded period by 0.5 seconds, I'm not sure if you'll see a significant difference, but you could, and then also, it just kind of makes the whole experience a bit longer. But it's generally as when people are in simulators or on the road and you ask them to perform a task, they don't always look away for one second or, you know, 1.5 seconds during the fast performance part."

Another respondent said:

> "What's confusing is the Alliance speaks of a 1.0 second shutter-closed occlusion time. I think this could be an opportunity to change this – to go to 1.5 seconds – this is a chance to manage all those after-image effects when I look to visual behavior. And to remove the confusion that the discrepancy causes. Also, looking at glances that are made back to the road from the display – this is not done in one second. The 1.5 seconds closed interval in ISO is corresponding more to natural behavior than the Alliance guideline. If you look at eye-tracking data – for example look to rearview mirror or glance to instrument reading it is about a 600 milliseconds glance and some saccade here (at beginning) and there (at the end) … traffic scenery… it is not possible to do in one second."

Another respondent said:

> "Of course it would, but the reason to have that value is to show that a longer glance is not needed to complete the task."

Another expert felt that it wouldn't make a significant difference.

In general, it seems that the experts did not feel that changing the occlusion time from 1.0 second to 1.5 seconds would make a significant difference in performance. However, one expert felt it would be more realistic to change the occlusion time as it would better reflect real driving behavior.

Is the 15 seconds TSOT excessive? Not sufficient? Why?

Two respondents felt the 15 second TSOT was reasonable. One stated that:

> "I think the 15 seconds TSOT is a reasonable compromise because it recognizes the need to allow some visual tasks, but the 1.5 second shutter open time restricts long glances and 10 intervals helps to avoid lengthy tasks."

Another replied:

> "Like the occlusion procedure itself, 15 seconds may not be the perfect value but it is a reasonable and useful value based on extensive testing and negotiation. As you are aware, there are some that think the driver is always responsible for safe control of the car including when and for how long to engage in secondary tasks. Therefore, any specific criterion that limits freedom of response takes away the driver's responsibility: this should not be specified. The 15 seconds criterion value was based on long, thoughtful negotiation as well as scientific evidence."

Another said that to reduce the TSOT would mean:

> "Someone would have to have some very strong [evidence indicating] why they need to reduce it because at this time, there are a lot of traditional tasks in vehicles that could

> take more, or could take up to, you know, the 15 seconds to perform and if that was reduced, someone would have to go through an exhaustive review of a lot of different vehicles and a lot of different tasks, including climate control, vehicle settings, infotainment, etc. and have a really good understanding as to what they are going to potentially be limiting."

It was also reported that:

> "Occlusion is not sensitive on short tasks. Tasks under 5 seconds or 7 seconds, (this is a real disclaimer) shouldn't be tested by occlusion."

Somewhat advocating an extension of the 15-second TSOT threshold, an expert said:

> "One has to look at 2 criteria: 1) TSOT and 2) the longer the TSOT gets, the more important interruptability is. When the dialog is long, it is important to know whether it can get further prolonged by interruptions. What we couldn't achieve was to get a rationale that says if a task is short – it will have only some glances – and one could assume it is interruptible. The longer [the task], the more it has to prove it is interruptible. I miss this [extra criterion of interruptability] in the Alliance guidelines. This would make things complicated because it isn't a single criterion, but I think one could have a formulation – where if you go beyond a given threshold you have to prove that the dialog is made for interruptions and there are subtasks. It is about the quality of the dialog. There was one study on this; I couldn't find it – an old publication from Christopher Monk – also on interruptability."

One respondent stated:

> "From my personal experience with the occlusion method I cannot actually give a reasonable assessment. However, because of results as, e.g., those reports by Baumann et al. (2004, Exp. 2), I am a little bit cautious with static time-based criteria because, from my point of view, chunkability and resumability are more important."

Another expert provided similar feedback, indicating that:

> "Occlusion is a method that takes a look at interruptability and resumability; therefore I think that the attention should be on the R metric as a criterion. In general, there is some doubt on how limiting any task time will lead to good results in terms of driver distraction as single tasks can be arbitrarily added by drivers. For example, we have one 15 seconds destination entry task and one 15 seconds radio task that can be easily combined to a 30 seconds task if performed successively. What is now the result for driver distraction? I believe that other criteria like interruptability are much more important."

In general, the experts agreed that the 15 TSOT was not excessive and seemed to be a good value to use. If the threshold was shortened, then there was some concern that much of the current instrumentation in vehicles would have to be reevaluated to verify that it meets a shorter TSOT.

Several experts used this question to emphasize that TSOT was not as important as task interruptability when looking at driver distraction in relation to performing tasks, stressing the importance of considering thresholds for interruptability in occlusion testing.

Do you consider the 15 seconds TSOT threshold to be equivalent to the 2/20 seconds thresholds? Why or why not?

Overall, half of the respondents felt there was a correlation between the 15 second TSOT threshold and the 2/20 second threshold. One expert replied:

> "Yes, it is correlated with 20 seconds. Also, the '2 second' glance corresponds to the shutter open time minus a saccade at the beginning. 20 seconds corresponds to 10 glances, or shutter openings. We don't need saccades [represented]."

Vision is suppressed during saccades – which is perhaps why the expert said saccades are not needed, although this reason was not specifically articulated.

Another expert offered this response:

> "I think the 15 second [criterion] is very much correlated with the 20 second [criterion]. If you have one, you get the other by doing correlations. You take 15 second or 20 second in natural data; those are correlated with total task time. I would supplement those with the interruptability metric. It makes it more robust."

A third respondent said:

> "Ample evidence was developed by Alliance members (and in SAE & ISO) to show that there is a strong correlation between TSOT and TGT with TSOT always being less than or equal to TGT. It is believed that there is an 'efficiency' developed in the occlusion technique (perhaps by mental rehearsal during the occluded period) that results in TSOT being less than TGT."

One expert felt there was a difference. He replied:

> "I think the 15 second TSOT is more stringent than the 2/20 thresholds because it limits the total eyes-off-road time (15 compared to 20) and the maximum glance duration is less (1.5 compared to 2), although the 2/20 threshold is under dynamic driving conditions, so the driver has fewer resources than is the case with the occlusion method. The fact that the 2/20 threshold allows 10 two-second glances is very worrying. I would feel extremely uncomfortable with that amount of visual inattention. Not to mention that 15% of glances[2] (using the Alliance 85th percentile criterion) can exceed

[2] This expert uses a mistaken application of the Alliance 85th percentile criterion. The criterion actually means that 85% of *participants* in the test sample must have a mean glance duration that is less than or equal to 2 seconds. Thus, only 15% of the participants in the test sample may have a mean glance duration that exceeds 2 seconds. **For a sample of 10 participants** (who must be aged 45 – 65 years under the Alliance test procedures), this would mean that if more than a single participant had a mean glance duration over 2 seconds, the task would <u>not</u> meet the criterion. Also, consider that in this older age range, age effects are already significant, making this assessment *quite different from* a similar criterion applied to a test sample of 20-year-olds.

the 2/20 threshold. So a 30 second glance away from the road would be acceptable as long the 85% of glances are within the 2/20 limit. The 2/20 numbers and the rationale for the 85 percentile are not supported by recent research evidence either. The premise for these numbers is debatable and it is based on very old data that pre-dated telematics devices."

The remaining experts either just explained what the two thresholds meant, or didn't answer the question. In general, there is a perception that the 15 second and 2/20 second thresholds are correlated.

Can you compare this to the 7.5 second JAMA-suggested threshold?

One expert indicated that:

> "It boils down to JAMA's definition of a task and ours. If you were to enter a destination by a phone number as a task, we would include everything from the first press of the NAV button to get you into NAV to selecting a destination to selection definition by phone number, then entering potentially the 10-digit phone number and then hitting set as destination. For JAMA this same scenario would be a handful of separate tasks. So entering NAV as a task, and it takes you up to 7.5 seconds to press a button, like in destination can take you up to 7.5 seconds, entering the area code takes you up to 7.5 seconds and then your prefix and numbers, 7.5 seconds and then hit [an execute key], 7.5 seconds. The difference is in how a task is broken down or interpreted. If we are to apply those 7.5 seconds to the way we define a task, you would find that drivers will be driven mad and we will be limiting things that we should not be limiting in the vehicle."

One expert felt the JAMA work was based on subjective utterances data which led to an unnecessarily strict criterion. He felt that the Alliance work was based on objective performance data and a better way to draw occlusion criteria. He also stated:

> "I don't see a breakdown in driving performance measures at 7.5 seconds."[3]

Finally one expert stated that from his personal experience with the occlusion method he couldn't give an opinion. In general, the experts were not particularly complimentary of the 7.5-second threshold suggested by JAMA.

---

[3] Interestingly, this expert was the only one to allude to a longstanding controversy over the findings in European vs. American and Japanese driving performance data. The U.S. and Japanese data show strong relationships between visual demand and lane keeping measures – whereas European data have shown much weaker relationships between visual demand and lane keeping -- at least to the extent that similar tasks could be compared under somewhat similar driving conditions. This fundamental difference in findings has contributed to controversy over measures of visual demand, and particularly the level at which a "redline" should be established. The reasons for the differing magnitudes of relationship between visual demand and lane keeping performance are not yet known – but there are several possibilities, which range from higher levels of skill and training among European drivers at licensing, to different (more tightly controlled) vehicle dynamics in the European vehicles, to different driving practices and behaviors between cultures, to differences in methodology between studies.

What would be appropriate (minimal) methods for statistical comparisons to provide reasonable statistical confidence and why?

Two respondents felt small sample sizes were appropriate to use. One stated:

> "Not being a statistician I think, from a strict statistical view, the larger the sample the better and a power analysis can be performed to derive the ideal sample size. However, the guidelines were developed to be practical, implementable and useful. I personally believe in the 'power of small numbers' (see Virzi, 1990 – Streamlining the Design Process: Running Fewer Subjects; Proceedings of the Human Factors Society). Virzi states that with 4 or 5 subjects 80% of the usability problems are detected and additional subjects are less and less likely to reveal new information. That is, while not meeting rigorous statistical requirements, a small sample will supply enough data to determine if an interface is likely to cause problems for the majority of users. Additional subjects help control for the (potentially large) variability among users but it quickly becomes prohibitive for a manufacturer to maintain large sample sizes for every evaluation. Of course, allowing a small number of 'ringers' to be used for the sample must be avoided as well."

Another replied:

> "Well the way we do it, is we just basically just count the open period that is required to perform the task. If you either do it with 10 or fewer or not and then you have to have 85 percent of your participants, using 10 or fewer glances or open periods. We don't get into partial open periods and things like that. So if they finish a task, say you know in a tenth of a second into the last open period, we don't slice it that finely."

One expert stated:

> "We are not really talking about which statistical methods to use here. Part of this has to do with how robust the method is and how replicable it is. We recently did an occlusion test on text editing and production. What astonished me was a student doing this test with the ISO document came to nearly the same values found by others. This for me again proves that the method is highly replicable, and objective, which, for statistics, is very nice. It is replicable, robust. A student did it who wasn't involved in any ISO history. The method has high power which means one can work with small sample sizes."

Two other respondents replied about what test to use and confidence issues. One said:

> "Any decision rule should be based on an alpha of 0.05, although this could be a 1-tailed test given that it is reasonable for the evaluators to have directional hypotheses. These decisions are safety-relevant and standards should be high. 95% confidence is conventional in the domain of Human Factors. There are also the issues of outliers, data treatment and transformation, non-parametric."

The other respondent simply stated:

> "T test for comparing occlusion results."

Finally an expert wrote:

> "I am not sure if the conventional way of testing for statistical significance is the appropriate approach because one tests against H0 [the null hypothesis]. This means that H0 ($\mu 1=\mu 2$) is rejected (and H1 is accepted) if the probability of the data given H0 is less than 5% (or whatever level was chosen). However, when testing against a certain reference criterion one might argue that it is more appropriate to do this the other way around, i.e., test against a H1 (μ > reference value) and accept a H0 (μ <= reference value) if the probability of the data given H1 is below a certain percentage level. This could be a little bit more conservative approach (in the sense of erroneously accepting a too demanding solution) but might reflect perhaps the actual problem a little bit better."

Overall, the experts approached the question somewhat differently. Some talked about the sample size (see also the next question). Others talked about hypothesis testing approaches. In general, the experts did not express any particular preference towards any particular statistical methods. There were also comments about procedural approaches to the test itself.

Statistically speaking, what should be considered a sufficient sample size for these (or any other suggested) metrics in order to establish compliance?

Answers to this question ranged from 10 to 20 participants. One respondent stated:

> "As a 'rule of thumb' I would consider N = 10 (valid cases) as a minimum sample size. But therefore all measures taken to control for error variance have to be kept in mind. But keep in mind that following the conventional way of testing for statistical significance the situation might become a little bit paradoxical because in a statistical sense compliance might translate to accepting H0. In this case, lower sample sizes could become recommendable because they reduce the power of the statistical test and thereby the probability of a statistically significant result (i.e., reject H0 and accept H1)."

Another expert replied:

> "It was shown that 10 -15 participants with 5 trials lead to a good statistical result."

Three respondents felt 20 participants was a good sample size. One said:

> "There needs to be a consideration of sensitivity, reliability and practicality. Of course the power of the test also needs to be considered and the magnitude of the differences between conditions. If you cornered me – I'd say N = 20 simply because any more than that becomes too impractical. Other issues like the diversity, age and experience of your sample need to be considered and reported because this variability will cause

> issues with smaller sample sizes. Whatever you do, be sure to report relevant information such as participant characteristics, dropouts, outliers, data treatment, etc."

Another of the three respondents suggesting 20 participants indicated that a sample of 10 participants was too small. He replied:

> "If applied correctly one could use a small sample size (in ISO it is small – 10 or something like that). I think that is not enough, but say 20 subjects. If training procedures are followed, and enough repetitions are obtained, then 20 subjects are plenty enough to do those tests. The method is robust and sensitive."

How many repetitions should be obtained per participant and why?

The responses ranged from two to five or perhaps more depending on the task. Two experts felt whatever the current standard is would be fine. One went on to say:

> "The crucial point is the training. Train subjects, stay with a balanced design, and number of repetitions. The figure that is too small in the standards from my perception is the number of subjects."

Another respondent felt three repetitions were necessary in order to get a more stable indication of performance. He felt repetitions would also help to show learning trends. One expert indicated:

> "As a 'rule of thumb' I would recommend four repetitions (without training trials) based on my personal experiences. The more interesting question would be how to deal with outliers and missing data."

One expert said:

> "Ideally the number of repetitions should be determined by performance to criterion. Depending on the subject and the task, sometimes one repetition is all that is needed for a very simple task, e.g., turning heat on, while for a more difficult task, e.g., destination entry by street address, perhaps 5 repetitions are insufficient for errorless performance."

Do you consider occlusion as an early test, a final verification test, or somewhere between these two extremes? Can you provide reasons why you place it where you do?

All respondents agreed that occlusion testing could be used early, once a suitable prototype is developed, and some responded that you could also use it for final verification. As one expert said:

> "It is especially useful in early stages to identify design options which are obviously poor and should not be considered by more sophisticated and costly procedures in later stages."

One respondent stated that not everything can be tested via occlusion simply because there are too many tasks to test. His organization narrows it down by only testing the new features that have been added to a device or system.

Another emphasized that:

> "[Occlusion] could be performed multiple times in design cycle. The effort is small. But it must be emphasized that it is simply restricted to visual demand (and readability/interruptability). Occlusion is a very sharply-tailored method. It doesn't give you measures of cognitive or motoric loads, etc."

One stated that:

> "Occlusion is most frequently applied (at least in my organization) as a verification test. In some cases this results in limiting functionality of an interface while moving. Such limiting would be to display only the first N items in a list and not the full list while moving. When stopped, the complete list could be reviewed by the driver."

Do you think occlusion tests should be performed multiple times in the design cycle? How often?

Two respondents replied yes. One said he didn't think it should be required to perform occlusion tests multiple times. Another stated:

> "After a design option has passed a more sophisticated test (e.g., in the simulator) occlusion tests may become redundant."

One also stated:

> "If there are questions about the interface accessibility while moving it can be applied at that time. If resources are available to support such testing it is good to do. However other techniques (such as counting the number of steps or estimation of total glance time or total task time) are more appropriate for design iterations, with occlusion being used for verification."

Is there anything that you would like to add to the discussion that we've had today? That is, is there some topic or issue that we did not include in this questionnaire that you feel might be valuable?

One expert responded:

> "Occlusion is very efficient. When you [talking to the interviewer] showed CAMP data and analyzed CAMP data, it corresponded highly with what we had found. It is highly replicable, – more than almost any other method in human factors.
>
> In the future, one issue that may need to be addressed is the equipment. There is only one supplier of occlusion goggles, – and within ISO we find that it can be a bottleneck,

waiting to get the occlusion goggles we need. We have to find alternative suppliers for the goggles, or alternatives to the goggles. Occlusion is applicable not only to vehicles but to portable devices and apps. What I see when I go to conferences is that the bigger problem is the mobile phones and apps market. Those guys need a feeling for visual demand. We gave it to colleagues in Computer Science department. They did occlusion tests and saw how visually loading their layout was. They didn't go into a simulator or field. They did a short test and the story was clear. Occlusion is very telling about visual demand, and very easy/quick to use."

One expert brought up the concept of simplifying models.

"There needs to be simplified models that are used and you have to have simplified models to provide guidance early in design work. Otherwise, you know at the 11th hour you'll perform some sort of verification and there could be tons of changes at that point, you know, limiting things, locking things out or making them available. Models where we can estimate per task the number of glances that will be used."

This expert went on to say that companies may be approaching hundreds of different tasks that could be evaluated and that it's not possible for anyone to use any method, such as occlusion, or even taking performance methods in a simulator or on the road to evaluate them.

"Testing all the different things in vehicle is impossible, even finding, training someone on all of those would be completely impossible. You know, I've worked on a lot of these systems and there are just so many different means and methods of doing things that for you to be fairly proficient in the guidelines to do a – you have basically a three month in-service proficiency level, you have to be training someone for hours or days on it. From my perspective, the main thing that we need to do is provide guidance early on when the systems are being developed so that they can be designed so that these limitations are inherent in the initial design, otherwise what happens is you design a system, you go through it and when you cut and chop it up and the overall usability of the entire system is degraded and I strongly believe that can cause more problems than if you left some things open that didn't meet the guidelines. A good example would be POIs for fast food restaurants. Let's say that displaying the list of potential candidates required 11 glances. If that was discovered using a confirmation prototype, where we're getting ready to lock things in for final release to production, they may have to cut it off at that point without changing the entire system, so you have people driving along thinking they can do it. When they get to that point the function is locked out. [Drivers] then have to backtrack and do all that.

So that's something that people need to understand when they're considering new guidelines. When you limit stuff you can do it in a good fashion where you don't have people searching and searching forever, but you can also screw things up where they have been going down a path with a desired goal and you stop them at a non-sensible location.

We have short timelines and you have to have tools where you can provide guidance early on, the verification for the entire system needs to be rather quick so that we don't have any of these crazy things where at the end, you know, things are cut and chopped and everything else because that, again, can cause as many or more problems than anything else."

SUMMARY

Overall, the experts expressed limited agreement in their responses, but some key themes emerged. These are summarized as follows:

- Experts seemed comfortable with using the occlusion method as a tool to assess visual demand, but preferably in the company of other methods.
- The experts tended to agree that the shutter open interval was reasonable and did not feel that changing the occlusion time from 1.0 to 1.5 seconds would make a substantial difference in performance.
- In general, the experts agreed that the 15 TSOT was not excessive and seemed to be a good value to use.
- In general, there was a perception that the 15 seconds and 2/20 seconds thresholds are correlated. Experts expressed concerns about the 7.5 seconds threshold suggested by JAMA.
- Experts generally believed that the number of participants in occlusion tests should range from 10 to 20 participants, performing from two to five repetitions per task.

ACKNOWLEDGMENTS

We gratefully acknowledge all the experts that participated in the interviews for their time and insights. These included Drs. Klaus Bengler, Peter Burns, James Foley, Christhard Gelau, and Andreas Keinath.

REFERENCES

Angell, L., Auflick, J., Austria, P. A., Kochhar, D., Tijerina, L., Biever, W., Diptiman, T., Hogsett, J., & Kiger, S. (2006). *Summary of Some Key Findings on Visual Occlusion from the CAMP Driver Workload Metrics Project. Driver Workload Metrics Task 2 Final Report & Appendices. Final Research Report*. (Report No. DOT HS 810 635). Washington, DC: National Highway Traffic Safety Administration.

Asoh, T. & Iihoshi, A. (2006). *Occlusion method to evaluate visual distraction to be caused by using car navigation systems*. Paper presented at Driver Workload Metrics Workshop, October 2-3, Ottawa, Canada.

Baumann, M., Keinath, A., Krems, J. F., & Bengler, K. (2004). Evaluation of in-vehicle HMI using occlusion techniques: Experimental results and practical implications. *Applied Ergonomics, 35*, pp. 197-205.

Brook-Carter, N., Stevens, A., Reed, N., & Thompson, S. (2009). Practical issues in the application of occlusion to measure visual demands imposed on drivers by in-vehicle tasks. *Ergonomics, 52*(2), pp. 177-186.

Driver Focus-Telematics Working Group. (2006). *Statement of principles, criteria and verification procedures on driver interactions with advanced in-vehicle information and communication systems* (p. 90). Washington, DC: Alliance of Automobile Manufacturers.

Foley, J. P. (2008). Now You See It, Now You Don't: Visual Occlusion as a Surrogate Distraction Measurement Technique. In M. A. Regan, J. D. Lee & K. L. Young (Eds.), *Driver Distraction: Theory, Effects, and Mitigation* (pp. 123-133).

Gelau, C., Henning, M. J., & Krems, J. F. (2009). On the reliability of the occlusion technique as a tool for the assessment of the HMI of in-vehicle information and communication systems. *Applied Ergonomics, 40*(2), pp. 181-184.

Gelau, C., & Schindhelm, R. (2010). Enhancing the occlusion technique as an assessment tool for driver visual distraction. *Intelligent Transport Systems, IET, 4*(4), pp. 346-355.

Green, P. A. (2002). Where do drivers look while driving (and how for long)? In R.E. Dewar and P.L. Olson, (Eds.), *Human factors in traffic safety*. Tucson, AZ: Lawyers and Judges, (pp. 77-110.)

Hashimoto, K. & Atsumi B. (2001). *Study of Occlusion Technique for Making the Static Evaluation Method of Visual Distraction.* Exploring the Occlusion Technique: Progress in Recent Research and Applications Workshop, Torino, Italy.

Horberry, T., Stevens, A., Burnett, G., Cotter, S., & Robbins, R. (2008). *Assessing the visual demand from in-vehicle information systems by means of the occlusion techniques: the effects of participant age*. Paper presented at the IET Intelligent Transport Systems.

International Standards Organization. (2007). *Road vehicles - Ergonomic aspects of transport information and control systems - Occlusion method to assess visual demand due to the use of in-vehicle systems* (International Standard No. ISO 16673). Geneva: International Standards Organization.

Japan Automotive Manufacturers Association (JAMA). (2003). *Guidelines for In-Vehicle Display Systems – Version 3.0*. Tokyo, Japan.

Karlesson, R. & Fichtenberg, N. (2001). *How different occlusion intervals affect total shutter open time, Reg. no. 202-0006*. Volvo Car Corporation, Exploring the Occlusion Technique: Progress in Recent Research and Applications Workshop, Torino, Italy.

Krems, J. F., Keinath, A., Baumann, M. Bengler, K. & Gelau, C. (2000). *Evaluating Visual Display Designs in Vehicles: Advantages and Disadvantage of the Occlusion Technique*. Chemnitz Technical University, Chemnitz, Germany.

Monk, C. A., & Kidd, D. G. (2007). *R We Fooling Ourselves: Does the Occlusion Technique Shortchange R Estimates?* Paper presented at the Proceedings for 4th International Symposium on Human Factors, Driver Assessment, Training, and Vehicle Design, Stevenson, Washington.

Noy, Y. I., Lemoine, T. L., Klachan, C., & Burns, P. C. (2004). Task interruptability and duration as measures of visual distraction. *Applied Ergonomics, 35*, pp. 207-213.

Society of Automotive Engineers. (2004). Standard of Practice, SAE J2364 200408, "*Navigation and route guidance function accessibility while driving*." Warrendale, PA: SAE International.

Senders, J. W., Kristofferson, A. B., Levison, W. H., Dietrich, C. W., and Ward, J. L. (1966). *An Investigation of Automobile Driver Information Processing* (*BBN Technical Report 1335*). Cambridge, MA: Bolt Beranek and Newman.

Senders, J. W., Kristofferson, A. B., Levison, W. H., Dietrich, C. W. and Ward, J. L. (1967a). *An Investigation of Automobile Driver Information Processing* (*BBN Technical Report 1482*). Cambridge, MA: Bolt Beranek and Newman.

Senders, J. W., Kristofferson, A. B., Levison, W. H., Dietrich, C. W., and Ward, J. L. (1967b). The Attentional Demand of Automobile Driving. *Highway Research Record #195*, pp. 15-33.

Senders, J. W. and Ward, J. L. (1968). *Additional Studies of Driver Information Processing* (*BBN Technical Report 1738*). Cambridge, MA: Bolt Beranek and Newman.

Stevens, A., Bygrave, S., Brook-Cater, N. & Luke, T. (2004). *Occlusion as a technique for measuring In-Vehicle Information System (IVIS) visual distraction: a research literature review*, TRL Report 609. Crowthorne: Transport Research Laboratory.

Stevens, A., Burnett, G., & Horberry, T. (2010). A reference level for assessing the acceptable visual demand of in-vehicle information systems. *Behaviour & Information Technology, 29*(5), pp. 527-540.

APPENDIX A. INTRODUCTORY E-MAIL FOR INITIAL CONTACTS WITH EXPERTS

Dear [*Name of Contact Person*],

The National Highway Traffic Safety Administration (NHTSA) of the U.S. Dept. of Transportation has initiated a research project with us, the Virginia Tech Transportation Institute, which is intended to explore certain issues related to occlusion and how it is employed in addressing driver distraction for in-vehicle devices. As a key part of this project, we are communicating with researchers in the field of distraction and in-vehicle device development to get their input, and we would like to include your perspective.

The areas to be addressed will include:

- The occlusion method as currently implemented in the Alliance guidelines
- Whether the occlusion metric should be considered for use in visual-manual distraction prevention guidelines, in light of
    - Recent research on the occlusion method
    - Inherent advantages and disadvantages of occlusion and traditional eye glance measures
- Whether adjustments to the occlusion metric, as it stands today, should be considered

If you are willing to give us your input on these topics, please detach the RSVP form which accompanies this invitation, fill it out, and return it to us by email within 10 days so that we may set up an appointment with you for a telephone interview. If you prefer, we can send you a written questionnaire to fill out as opposed to doing a telephone interview. In order to answer any questions that you may have about this effort, and in order to complete scheduling of an interview if applicable, we may follow up with you by phone within the next week, after you have had a chance to reflect on this and return the attached short form. Participation in this interview/questionnaire is entirely voluntary.

If you choose to participate in a telephone interview for your planning purposes, you will receive a short packet prior to the telephone interview (which includes some reference material for use during the interview). We ask you to have it with you when we hold the telephone interview. The telephone discussion will last about one hour – and will be recorded (if you grant permission) – otherwise, we will only take written notes of your answers.

If you prefer to respond by a written questionnaire, we will send you the full questionnaire, which will have embedded in it the reference materials. Providing answers to the questionnaire should take you about 1 to 2 hours to complete.

The intent of gathering this information is to gain your perspective on these important topics, and we assure you that we will protect the confidentiality and security of the information you provide. We will summarize your answers with those of other experts, in such a way that no particular answer can be traced to any particular expert by anyone outside the research team. Only the VTTI portion of the research team will have access to the interview recordings, written questionnaires, and notes. The NHTSA portion of the research team will have access to aggregated data (pooled and summarized across interviews). During the interview, or at the end of the written questionnaire, we will ask you if you would like to be identified as a contributor to the research effort in any report or publication that should emanate from this work (with the provision that your name will not be associated with any particular response). You will be listed as a contributor only if you wish to be.

Participation in this interview process is entirely voluntary, as is filling out the written questionnaire. You are free to opt out, to decline to answer any question at any time, and to withdraw from the interview at any time, without any penalty or consequence. The same is true for deciding not to fill out the written questionnaire. This research study has been reviewed by the Virginia Tech Institutional Review Board. For research-related problems or questions regarding your rights as a research participant, you may contact Dr. David Moore, Chair, Virginia Tech Institutional Review Board for the Protection of Human Subjects, telephone (540) 231-4991; e-mail: moored@vt.edu; address: Office of Research Compliance, 2000 Kraft Drive, Suite 2000 (0497), Blacksburg, VA 24060. For other questions, please feel free to contact Miguel Perez or Linda Angell. Thank you for your help with this important effort; we look forward to speaking with you soon!

Sincerely,

Miguel Perez, Ph.D. & Linda S. Angell, Ph.D.
Human Factors Research Scientists
Center for Automotive Safety Research
Virginia Tech Transportation Institute
Blacksburg, VA 24061
540 -231-1537 -----------------313-881-6641

Form for RSVP to Interview Invitation

Please fill in the information below and **return it by email to us at: miperez@vtti.vt.edu**.

Your Name: ____

1. I would like to participate in an interview discussion (written or over the phone) on these topics.

   ___ Yes, written responses to a list of questions

   ___ Yes, over-the-phone interview

   ___ No

   If no, could you suggest another expert that may be able to participate in an interview? ____

   (If no, we thank you for your time, please skip the remainder of the form and return it to us)

2. *(If you choose to have an over-the-phone discussion)* I consent to having the over-the-phone discussion-session tape-recorded, given that the information I provide (and the recording of it) will be protected as confidential information, will not be shared in identifiable form with anyone outside the VTTI research team, except as part of an aggregated pool of data (to which NHTSA will have access), and will be stored securely.

   ____ Yes, I am comfortable having the phone-discussion recorded.
   ____ No, I prefer to discuss the topics without being recorded. I understand that will mean that interviewers will take notes on the discussion – and I agree to allow notes to be taken. I understand that these interview notes will be protected as confidential information, will not be shared with anyone outside the VTTI research team, except as part of an aggregated pool of data (to which NHTSA will have access), and will be stored securely.

3. I may be contacted at the following phone number to set up an interview appointment:

   ____

Thank you!! We will contact you within a week, if you have indicated agreement to participate.

Should you have any questions about this, you may call: Dr. Miguel Perez (of Virginia Tech – at 540-231-1537). **REMEMBER**: Please return this RSVP form **within 10 days** to: **miperez@vtti.vt.edu**.

APPENDIX B. PACKET SENT TO EXPERTS PERFORMING PHONE INTERVIEWS

Table of Recent Research on the Occlusion Method

Source	Summary
ISO (2007)	**Vision and occlusion intervals** The vision interval is defined as 1.5 seconds and the occlusion interval is 1.5 seconds, which are considered to be consistent with the occlusion literature. **Task timing** The total shutter open time (TSOT) is defined as follows: –Start: Timing starts with the beginning of the first vision interval. –End: Timing ends when the instructed task has been completed and the participant says he or she is "done." –Duration: Tasks are timed from start to end without interruption, including errors and subtracting occlusion intervals. Individual system response delays greater than 1.5 seconds are accounted for by a procedure in the standard's appendix. If the task is completed during a vision interval, then only that part of the vision interval that was used for the task should be included in the TSOT. As an alternative, TSOT may be approximated by the number of vision intervals needed to complete the task multiplied by the 1.5 second vision interval. Another approximation is provided by (TTTOccl /3.0) 1.5. Additional details are provided in the standard for the calculation of total task time unoccluded (TTTUnoccl) and total task time in occluded conditions (TTTOccl).
Alliance of Automobile Manufacturers (2003)	The Alliance guidelines state that TSOT, for a task accessible to the driver while driving, should not exceed 15 seconds.
Japanese Automobile Manufacturers Association (2003)	The JAMA guidelines (Section 4.2) state that 'The visual information to be displayed shall be sufficiently small in volume to enable the driver to comprehend it in a short time or shall be presented in portions for the driver to scan them in two or more steps' (p.2). The occlusion method is included in the verification techniques in Annex 3 to the JAMA guidelines, stating that the total shutter opening time (with vision intervals of 1.5 seconds and occluded internals of 1.0s) should be no more than 7.5 seconds. This criterion is based on the observed correlation between TSOT and total glance time (TGT), measured using direct eye glance performance (Karlesson and Fichtenberg, 2001, Hashimoto and Atsumi, 2001).
van der Horst (2004)	Provides a summary of work conducted using PLATO occlusion goggles. The author concludes from his summary that visual occlusion is an effective tool for addressing visual sampling behavior and workload while driving. It is also an effective tool to assess the safety effects of in-vehicle devices.

Baumann, Keinath, Krems and Bengler (2004)	Concluded, based on their empirical data, that the occlusion technique was a reliable and valid method for evaluating visual as well as dialogue aspects for in-vehicle information systems. In general, their results showed that the occlusion approach was, in terms of total task time and number of errors, more representative of secondary task performance under actual driving than static task performance would be.
Noy, Lemoine, Klachan, and Burns (2004)	Examined task interruptability using the occlusion technique. Based on their results, the authors could not suggest the use of TSOT or R value over the other. Tradeoffs between the measures exist. The authors do suggest, however, that both R and TSOT are needed, "perhaps even a composite measure such as $TSOT^2$/TTTunocc" (p. 212)
Asoh and Iihoshi (2006)	Presented data from various experiments designed to understand the relationship between TGT and other task assessment tools, including the occlusion method. They first measured TGT, lane deviation, and participant uneasiness for performing different navigation-related tasks. They compared TGT while performing tasks for different roadway types (2-lane urban, 1-lane urban, Joban Expressway, Metropolitan Expressway) and IVIS input devices (Touch panel, Joy Stick, Remote Control, Rotate). The authors found that the lowest TGT where 1) participants had neutral feelings concerning uneasiness, and 2) lateral control was not affected, was 7.9 seconds. Further findings reported by Asoh and Iihoshi (2006) suggested that TSOT had a stronger correlation with TGT (0.8933) than TTT (0.8760) for bench testing. The bench testing intervals for closed and open occlusion times in seconds were closed 0.5/opened 1.0, closed 1.0/opened 1.0, closed 0.5/opened 1.5, closed 1.0/opened 1.5, and closed 3.0/opened 1.5, with the closed 1.0/opened 1.5 alternative exhibiting the highest correlation. Based on these data, a TSOT of 7.5 seconds was found to be equivalent to 8 seconds TGT, which in turn was the TGT that resulted in neutral uneasiness levels.
Angell, Auflick, Austria, Kochhar, Tijerina, Biever, Diptiman, Hogsett, and Kiger (2006)	Conducted a review of various rules proposed for whether tasks were safe to complete while driving. They first interviewed experts in the field of driving and distraction as to whether certain task types were felt to need lower or higher levels of attention. They then ran a study having participants perform those tasks while in a static environment and recorded the time needed to complete tasks when vision was occluded and also when no occlusion goggles were used. Angell et al. (2006) compared the findings to what the experts agreed upon. The authors found that the best correspondence between different task classification rules and the expert's opinion was for a rule where a higher workload category was assigned to tasks whose TSOT exceeded 7.5 seconds. Angell et al. (2006) also showed correlations higher than 0.8 between TSOT and TGT.

Monk and Kidd (2007)	The authors conclude that the R value may be overestimated when participants do not have a tracking task to complete during the occlusion interval. It is also reported that "estimations of visual demand based on individual post-occlusion resumption times may provide a more precise measure of transition costs and resumability than measures based on Total Shutter Open Time." (p. 2). They propose an alternative formula for evaluating visual demand costs, using measurements of resumption lags and the times between unoccluded task inputs during each unoccluded episode. For their study, R and the proposed measures were not highly correlated. The authors state that this might be due to the small sample size of eight participants. They suggest that D may provide a more precise estimate of resumability because it is based on resumption measures.
Horberry, Stevens, Burnett, Cotter, and Robbins (2008)	State that guidelines should be developed for a detailed measurement protocol using standard occlusion. They specifically addressed the issue of age in relation to In-Vehicle Information System (IVIS) tasks. Their results showed differences in age with respect to TSOT and R. The older participants had more varied scores, especially for TSOT, and generally did more poorly than younger participants. The authors therefore suggest that to "obtain minimal inter-subject variability measurements [one] should use younger/middle-aged participants (broadly 17-66 years old)" (p. 175). The authors emphasize that this range merely reduces variability during testing and that the recommendation should not be taken as a suggestion that IVIS systems are developed only for younger drivers.
Gelau, Henning and Krems (2009)	Looked at how reliable the occlusion technique is for assessing visual distraction with IVIS. They looked at 4 different experiments that were previously conducted and looked at the internal consistency based on the TSOT. Their findings showed that "characteristic values for internal consistency were almost all in the range of .90 for the occlusion method and can be considered satisfactory."
Brook-Carter, Stevens, Reed, and Thompson (2009)	Looked at issues related to research gaps on system response delay, participant training levels prior to a study, number of repetitions for each task, sample size required, and calculating resumability. The authors suggest that three or more training tasks should be used for the more complex tasks being tested (e.g., place of interest search). Tasks that are less complex would require fewer training sessions. Their results also showed no significant difference with respect to the occluded versus non-occluded training method. Therefore, training with or without occlusion goggles shouldn't have a difference in the outcome of the research findings. In addition, based on differences between task completion times for successive trials of the same task, the authors recommend the use of 2 trials. In terms of sample size, the results suggested the use of over 10 participants for some tasks, although other tasks would benefit from a minimum sample size of 14 participants. The authors suggest a potential tradeoff of increasing the number of test trials rather than the sample size, which could be less time-consuming and expensive. Related to statistics, they suggest that "the value of R should be calculated on a within participant basis by taking the ratio of the respective median TSOT and TTTstatic across trial repetitions completed by one participant on each task under evaluation. Also, for the purpose of identifying interface designs that exhibit poor resumability, the 85th percentile value is suggested as the most appropriate"(p. 185). The use of the 85th percentile reduces the influence of outliers on the data.

Gelau and Schindhelm (2010)	Conducted a study having participants do a simple auditory tracking task during the occlusion period. Their findings supported those of Monk and Kidd (2007), discussed earlier. There was a significant increase in the TSOT and R value when participants performed the auditory tracking task. TSOT and R value measures obtained under an auditory tracking task also showed a clearer discrimination between an "easy" versus a "difficult" IVIS task than those obtained under non-auditory tracking task conditions. Furthermore, the R value was found to only be sensitive to the level of difficulty for the IVIS task when the auditory tracking task was used. The authors propose amending the standard occlusion method by having it include a cognitive load task, which they believe would improve the sensitivity of the method.
Stevens, Burnett, and Horberry (2010)	Point out that the current metric for the occlusion method lacks validation and does not "constrain measurement variability as much as it could" (p. 530). The authors believe the ISO standard is a good starting point but does not include all the measurement details. They proposed a new metric to test IVIS tasks that does not use the R value: Mean TSOT + (85th percentile – Mean)^2/Mean < 8.0s They justify the use of the 85th percentile because occlusion data are not usually normally distributed but instead follow a log-normal curve. Stevens et al. (2010) state that using the 85th percentile allows for more representation among participants who had more difficulty with some tasks. The use of a <8.9s TSOT is suggested as a conservative measure, and is based on JAMA's mean value of 7.5s. The authors adjusted that number to 8.0s to account for both the mean and the spread of the data.

APPENDIX C. SCRIPT FOR TELEPHONE INTERVIEWS WITH EXPERTS

Subtask 2c: Expert Review of the Visual Occlusion Method and How it Compares to Driver Eye-Glance Behavior

Script for Telephone Interviews with Experts

Hello – This is [names here] from the Virginia Tech Transportation Institute. Thanks for agreeing to speak with us today. We'd like to set the stage for our discussion today by reminding you that the National Highway Traffic Safety Administration (NHTSA) of the U.S. Dept. of Transportation has initiated a research project with us to explore issues surrounding the occlusion method in the context of use for guidelines to limit and test distraction from use of devices with visual-manual interfaces. As a key part of this project, we are speaking with researchers in the field to get their input, and we would like to include your perspective.

The areas we will discuss are:

- The occlusion method as currently implemented in the Alliance guidelines
- Whether the occlusion metric should be considered for use in visual-manual distraction prevention guidelines, in light of
  - Recent research on the occlusion method
  - Inherent advantages and disadvantages of occlusion and traditional eye glance measures
- Whether adjustments to the occlusion metric, as it stands today, should be considered

Before we get started, we would like to remind you that :

[Circle whichever selection, "a" or "b," that the participant made in his/her initial RSVP to participate and say that one over the phone:

a. "we will be recording this conversation just to make sure we catch everything. All recordings will be treated as privileged information, stored securely, and shared with no one outside the research team." or

b. "will not be recording this conversation in accordance with your wishes, so we will be taking notes as we discuss things with you. These notes will be treated as privileged information, stored securely, and shared with no one outside the research team."]

We expect this conversation to take about one hour.

If you feel uncomfortable answering any or all of the questions, you may opt out (or decline to answer) any questions at any time.

This research study has been reviewed by the Institutional Review Board at Virginia Tech. For research-related problems or questions regarding your rights as a research participant, you can contact: Dr. David Moore, Chair, Virginia Tech Institutional Review Board for the Protection of Human Subjects, telephone (540) 231-4991; e-mail: moored@vt.edu; address: Office of Research Compliance, 2000 Kraft Drive, Suite 2000 (0497), Blacksburg, VA 24060.

Now we would like to begin the interview.

I. BACKGROUND OF EXPERT

A. Please tell me a little bit about your experience and expertise in working in the area of driver distraction. [Probe for current job responsibilities, educational background/degrees, years of experience.]

B. (If not mentioned above), Do you also have experience and expertise in **working specifically** on the design and/or application of the occlusion method?

i. *No*

ii. *Yes* [If not mentioned in a.] Can you expand on the type of work that you've done with the occlusion method?

II. DISCUSSION OF CURRENT OCCLUSION TEST IN THE ALLIANCE GUIDELINES

Principle 2.1 of the Alliance Guidelines allows for the use of Visual Occlusion as a test method for verification using Alternative A. To pass, a static task should be completed with a total shutter open time ≤ 15 seconds (with reasonable statistical confidence). An equivalency between the 2 / 20 seconds Alliance thresholds and a 15 seconds Total Shutter Open Time (TSOT) is asserted to justify the use of the occlusion method. As described in the Alliance guidelines, the method uses 1.0 second occlusion periods with 1.5 seconds intervals of open shutter. The 15 seconds TSOT should be exceeded with reasonable statistical confidence.

A. Are you familiar with the Alliance treatment of the occlusion method?

B. Do you have any questions or want any additional explanation of how the Alliance implements the occlusion test?

Please keep this implementation in mind as the interview proceeds.

III. DISCUSSION OF RECENT RESEARCH

Research on the occlusion method has continued after the Alliance guidelines were published. The table on the next page, which has been provided to you electronically, contains a summary of that research. Please consider that research as I ask you the following questions.

Do you have that table handy for reference?

Yes/No

When “yes” – Okay, great! Let’s proceed.

Table of Recent Research on the Occlusion Method

Source	Summary
ISO (2007)	**Vision and occlusion intervals** The vision interval is defined as 1.5 seconds and the occlusion interval is 1.5 seconds, which are considered to be consistent with the occlusion literature. **Task timing** The total shutter open time (TSOT) is defined as follows: –Start: Timing starts with the beginning of the first vision interval. –End: Timing ends when the instructed task has been completed and the participant says he or she is "done." –Duration: Tasks are timed from start to end without interruption, including errors and subtracting occlusion intervals. Individual system response delays greater than 1.5 seconds are accounted for by a procedure in the standard's appendix. If the task is completed during a vision interval, then only that part of the vision interval that was used for the task should be included in the TSOT. As an alternative, TSOT may be approximated by the number of vision intervals needed to complete the task multiplied by the 1.5 second vision interval. Another approximation is provided by (TTTOccl /3.0) 1.5. Additional details are provided in the standard for the calculation of total task time unoccluded (TTTUnoccl) and total task time in occluded conditions (TTTOccl).
Alliance of Automobile Manufacturers (2003)	The Alliance guidelines state that TSOT, for a task accessible to the driver while driving, should not exceed 15 seconds.
Japanese Automobile Manufacturers Association (2003)	The JAMA guidelines (Section 4.2) state that 'The visual information to be displayed shall be sufficiently small in volume to enable the driver to comprehend it in a short time or shall be presented in portions for the driver to scan them in two or more steps' (p.2). The occlusion method is included in the verification techniques in Annex 3 to the JAMA guidelines, stating that the total shutter opening time (with vision intervals of 1.5 seconds and occluded internals of 1.0s) should be no more than 7.5 seconds. This criterion is based on the observed correlation between TSOT and total glance time (TGT), measured using direct eye glance performance (Karlesson and Fichtenberg, 2001, Hashimoto and Atsumi, 2001).
van der Horst (2004)	Provides a summary of work conducted using PLATO occlusion goggles. The author concludes from his summary that visual occlusion is an effective tool for addressing visual sampling behavior and workload while driving. It is also an effective tool to assess the safety effects of in-vehicle devices.

Baumann, Keinath, Krems and Bengler (2004)	Concluded, based on their empirical data, that the occlusion technique was a reliable and valid method for evaluating visual as well as dialogue aspects for in-vehicle information systems. In general, their results showed that the occlusion approach was, in terms of total task time and number of errors, more representative of secondary task performance under actual driving than static task performance would be.
Noy, Lemoine, Klachan, and Burns (2004)	Examined task interruptability using the occlusion technique. Based on their results, the authors could not suggest the use of TSOT or R value over the other. Tradeoffs between the measures exist. The authors do suggest, however, that both R and TSOT are needed, "perhaps even a composite measure such as $TSOT^2$/TTTunocc" (p. 212)
Asoh and Iihoshi (2006)	Presented data from various experiments designed to understand the relationship between TGT and other task assessment tools, including the occlusion method. They first measured TGT, lane deviation, and participant uneasiness for performing different navigation-related tasks. They compared TGT while performing tasks for different roadway types (2-lane urban, 1-lane urban, Joban Expressway, Metropolitan Expressway) and IVIS input devices (Touch panel, Joy Stick, Remote Control, Rotate). The authors found that the lowest TGT where 1) participants had neutral feelings concerning uneasiness, and 2) lateral control was not affected, was 7.9 seconds. Further findings reported by Asoh and Iihoshi (2006) suggested that TSOT had a stronger correlation with TGT (0.8933) than TTT (0.8760) for bench testing. The bench testing intervals for closed and open occlusion times in seconds were closed 0.5/opened 1.0, closed 1.0/opened 1.0, closed 0.5/opened 1.5, closed 1.0/opened 1.5, and closed 3.0/opened 1.5, with the closed 1.0/opened 1.5 alternative exhibiting the highest correlation. Based on these data, a TSOT of 7.5 seconds was found to be equivalent to 8 seconds TGT, which in turn was the TGT that resulted in neutral uneasiness levels.
Angell, Auflick, Austria, Kochhar, Tijerina, Biever, Diptiman, Hogsett, and Kiger (2006)	Conducted a review of various rules proposed for whether tasks were safe to complete while driving. They first interviewed experts in the field of driving and distraction as to whether certain task types were felt to need lower or higher levels of attention. They then ran a study having participants perform those tasks while in a static environment and recorded the time needed to complete tasks when vision was occluded and also when no occlusion goggles were used. Angell et al. (2006) compared the findings to what the experts agreed upon. The authors found that the best correspondence between different task classification rules and the expert's opinion was for a rule where a higher workload category was assigned to tasks whose TSOT exceeded 7.5 seconds. Angell et al. (2006) also showed correlations higher than 0.8 between TSOT and TGT.
Monk and Kidd (2007)	The authors conclude that the R value may be overestimated when participants do not have a tracking task to complete during the occlusion interval. It is also reported that "estimations of visual demand based on individual post-occlusion resumption times may provide a more precise measure of transition costs and resumability than measures based on Total Shutter Open Time." (p. 2). They propose an alternative formula for evaluating visual demand costs, using measurements of resumption lags and the times between unoccluded task inputs during each unoccluded episode. For their study, R and the proposed measures were not highly correlated. The authors state that this might be due to the small sample size of eight participants. They suggest that D may provide a more precise estimate of resumability because it is based on resumption measures.

Horberry, Stevens, Burnett, Cotter, and Robbins (2008)	State that guidelines should be developed for a detailed measurement protocol using standard occlusion. They specifically addressed the issue of age in relation to In-Vehicle Information System (IVIS) tasks. Their results showed differences in age with respect to TSOT and R. The older participants had more varied scores, especially for TSOT, and generally did more poorly than younger participants. The authors therefore suggest that to "obtain minimal inter-subject variability measurements [one] should use younger/middle-aged participants (broadly 17-66 years old)" (p. 175). The authors emphasize that this range merely reduces variability during testing and that the recommendation should not be taken as a suggestion that IVIS systems are developed only for younger drivers.
Gelau, Henning and Krems (2009)	Looked at how reliable the occlusion technique is for assessing visual distraction with IVIS. They looked at 4 different experiments that were previously conducted and looked at the internal consistency based on the TSOT. Their findings showed that "characteristic values for internal consistency were almost all in the range of .90 for the occlusion method and can be considered satisfactory."
Brook-Carter, Stevens, Reed, and Thompson (2009)	Looked at issues related to research gaps on system response delay, participant training levels prior to a study, number of repetitions for each task, sample size required, and calculating resumability. The authors suggest that three or more training tasks should be used for the more complex tasks being tested (e.g., place of interest search). Tasks that are less complex would require fewer training sessions. Their results also showed no significant difference with respect to the occluded versus non-occluded training method. Therefore, training with or without occlusion goggles shouldn't have a difference in the outcome of the research findings. In addition, based on differences between task completion times for successive trials of the same task, the authors recommend the use of 2 trials. In terms of sample size, the results suggested the use of over 10 participants for some tasks, although other tasks would benefit from a minimum sample size of 14 participants. The authors suggest a potential tradeoff of increasing the number of test trials rather than the sample size, which could be less time-consuming and expensive. Related to statistics, they suggest that "the value of R should be calculated on a within participant basis by taking the ratio of the respective median TSOT and TTTstatic across trial repetitions completed by one participant on each task under evaluation. Also, for the purpose of identifying interface designs that exhibit poor resumability, the 85th percentile value is suggested as the most appropriate"(p. 185). The use of the 85th percentile reduces the influence of outliers on the data.
Gelau and Schindhelm (2010)	Conducted a study having participants do a simple auditory tracking task during the occlusion period. Their findings supported those of Monk and Kidd (2007), discussed earlier. There was a significant increase in the TSOT and R value when participants performed the auditory tracking task. TSOT and R value measures obtained under an auditory tracking task also showed a clearer discrimination between an "easy" versus a "difficult" IVIS task than those obtained under non-auditory tracking task conditions. Furthermore, the R value was found to only be sensitive to the level of difficulty for the IVIS task when the auditory tracking task was used. The authors propose amending the standard occlusion method by having it include a cognitive load task, which they believe would improve the sensitivity of the method.

Stevens, Burnett, and Horberry (2010)	Point out that the current metric for the occlusion method lacks validation and does not "constrain measurement variability as much as it could" (p. 530). The authors believe the ISO standard is a good starting point but does not include all the measurement details. They proposed a new metric to test IVIS tasks that does not use the R value: Mean TSOT + (85th percentile – Mean)^2/Mean < 8.0s They justify the use of the 85th percentile because occlusion data are not usually normally distributed but instead follow a log-normal curve. Stevens et al. (2010) state that using the 85th percentile allows for more representation among participants who had more difficulty with some tasks. The use of a <8.9s TSOT is suggested as a conservative measure, and is based on JAMA's mean value of 7.5s. The authors adjusted that number to 8.0s to account for both the mean and the spread of the data.

A. In light of these findings and your personal experience, do you think these occlusion metrics are sufficient to describe visual demand for verification purposes? Why or why not?

B. Should the parameters for the occlusion test be adjusted?
   i. Could there be an influence from limiting the maximum glance duration at 1.5 seconds?
   ii. Could there be an influence from adjusting the occluded interval from 1.0 second to 1.5 seconds?

C. Is the 15 seconds TSOT excessive? Not sufficient? Why?
   i. Do you consider the 15 seconds TSOT threshold to be equivalent to the 2 / 20 seconds thresholds? Why or why not?
   ii. Can you compare this to the 7.5 seconds JAMA-suggested threshold?

D. What would be appropriate (minimal) methods for statistical comparisons to provide reasonable statistical confidence?

E. Statistically speaking, what should be considered a sufficient sample size for these (or any other suggested) metrics in order to establish compliance?
   i. How many repetitions should be obtained per participant?

F. Do you consider occlusion as an early test, a final verification test, or somewhere between these two extremes? Can you provide reasons why you place it where you do?
   i. Do you think occlusion tests should be performed multiple times in the design cycle? How often?

V. CLOSING QUESTIONS

Is there anything that you would like to add to the discussion that we've had today? That is, is there some topic or issue that we did not touch on during the interview that you feel might be valuable?
Okay. . . thank you! That completes the interview.
May we just ask you one final thing: Would you like to have your name included as a contributor to this research in any report or publication that may emanate from this work? (Please note that if we were to list your name, it would not be identified with any particular response or responses – rather, the list of contributors would be separated from the aggregated data and provided as part of an acknowledgement.)

RESPONSE: YES: ___ NO: ___

Thank you *so very much* for your time and your contributions!

APPENDIX D. WRITTEN QUESTIONNAIRE FOR EXPERTS

Subtask 2c: Expert Review of the Visual Occlusion Method and How it Compares to Driver Eye-Glance Behavior

Dear: [participant's name goes here]

Thanks for agreeing to give us your opinions on the visual occlusion method by filling out this questionnaire. I would like to remind you that the National Highway Traffic Safety Administration (NHTSA) of the U.S. Department of Transportation has initiated a research project with us to explore issues surrounding the occlusion method in the context of use for guidelines to limit and test distraction from use of devices with visual-manual interfaces. As a key part of this project, we are communicating with researchers in the field to get their input, and we would like to include your perspective.

The areas we will discuss are:

- The occlusion method as currently implemented in the Alliance guidelines
- Whether the occlusion metric should be considered for use in visual-manual distraction prevention guidelines, in light of
    - Recent research on the occlusion method
    - Inherent advantages and disadvantages of occlusion and traditional eye glance measures
- Whether adjustments to the occlusion metric, as it stands today, should be considered

Before we get started, we would like to remind you that filling out this questionnaire should take you approximately one to two hours.

If you feel uncomfortable answering any or all of the questions, you may opt out (or decline to answer) any questions at any time.

This research study has been reviewed by the Institutional Review Board at Virginia Tech. For research-related problems or questions regarding your rights as a research participant, you can contact: Dr. David Moore, Chair, Virginia Tech Institutional Review Board for the Protection of Human Subjects, telephone (540) 231-4991; e-mail: moored@vt.edu; address: Office of Research Compliance, 2000 Kraft Drive, Suite 2000 (0497), Blacksburg, VA 24060.

IV. BACKGROUND OF EXPERT

C. Please tell me a little bit about your experience and expertise in working in the area of driver distraction. Please include your current job responsibilities, educational background/degrees, and years of experience. Also include your experience and expertise in **working specifically** on the design and/or application of the occlusion method?

V. DISCUSSION OF CURRENT OCCLUSION TEST IN THE ALLIANCE GUIDELINES

Principle 2.1 of the Alliance Guidelines allows for the use of Visual Occlusion as a test method for verification using Alternative A. To pass, a static task should be completed with a total shutter open time ≤ 15 seconds (with reasonable statistical confidence). An equivalency between the 2 / 20 seconds Alliance thresholds and 15 seconds Total Shutter Open Time (TSOT) is asserted to justify the use of the occlusion method. As described in the Alliance guidelines, the method uses 1.0 second occlusion periods with 1.5 seconds intervals of open shutter. The 15 seconds TSOT should be exceeded with reasonable statistical confidence.

C. Are you familiar with the Alliance treatment of the occlusion method?

Please keep this implementation in mind as you answer the following questions.

VI. DISCUSSION OF RECENT RESEARCH

Research on the occlusion method has continued after the Alliance guidelines were published. The table on the next page, which has been provided to you electronically, contains a summary of that research. Please consider that research for the following questions.

Table of Recent Research on the Occlusion Method

Source	Summary
ISO (2007)	**Vision and occlusion intervals** The vision interval is defined as 1.5 seconds and the occlusion interval is 1.5 seconds, which are considered to be consistent with the occlusion literature. **Task timing** The total shutter open time (TSOT) is defined as follows: –Start: Timing starts with the beginning of the first vision interval. –End: Timing ends when the instructed task has been completed and the participant says he or she is "done." –Duration: Tasks are timed from start to end without interruption, including errors and subtracting occlusion intervals. Individual system response delays greater than 1.5 seconds are accounted for by a procedure in the standard's appendix. If the task is completed during a vision interval, then only that part of the vision interval that was used for the task should be included in the TSOT. As an alternative, TSOT may be approximated by the number of vision intervals needed to complete the task multiplied by the 1.5 second vision interval. Another approximation is provided by (TTTOccl /3.0) 1.5. Additional details are provided in the standard for the calculation of total task time unoccluded (TTTUnoccl) and total task time in occluded conditions (TTTOccl).
Alliance of Automobile Manufacturers (2003)	The Alliance guidelines state that TSOT, for a task accessible to the driver while driving, should not exceed 15 seconds.
Japanese Automobile Manufacturers Association (2003)	The JAMA guidelines (Section 4.2) state that 'The visual information to be displayed shall be sufficiently small in volume to enable the driver to comprehend it in a short time or shall be presented in portions for the driver to scan them in two or more steps' (p.2). The occlusion method is included in the verification techniques in Annex 3 to the JAMA guidelines, stating that the total shutter opening time (with vision intervals of 1.5 seconds and occluded internals of 1.0s) should be no more than 7.5 seconds. This criterion is based on the observed correlation between TSOT and total glance time (TGT), measured using direct eye glance performance (Karlesson and Fichtenberg, 2001, Hashimoto and Atsumi, 2001).
van der Horst (2004)	Provides a summary of work conducted using PLATO occlusion goggles. The author concludes from his summary that visual occlusion is an effective tool for addressing visual sampling behavior and workload while driving. It is also an effective tool to assess the safety effects of in-vehicle devices.
Baumann, Keinath, Krems and Bengler (2004)	Concluded, based on their empirical data, that the occlusion technique was a reliable and valid method for evaluating visual as well as dialogue aspects for in-vehicle information systems. In general, their results showed that the occlusion approach was, in terms of total task time and number of errors, more representative of secondary task performance under actual driving than static

	task performance would be.
Noy, Lemoine, Klachan, and Burns (2004)	Examined task interruptability using the occlusion technique. Based on their results, the authors could not suggest the use of TSOT or R value over the other. Tradeoffs between the measures exist. The authors do suggest, however, that both R and TSOT are needed, "perhaps even a composite measure such as $TSOT^2$/TTTunocc" (p. 212)
Asoh and Iihoshi (2006)	Presented data from various experiments designed to understand the relationship between TGT and other task assessment tools, including the occlusion method. They first measured TGT, lane deviation, and participant uneasiness for performing different navigation-related tasks. They compared TGT while performing tasks for different roadway types (2-lane urban, 1-lane urban, Joban Expressway, Metropolitan Expressway) and IVIS input devices (Touch panel, Joy Stick, Remote Control, Rotate). The authors found that the lowest TGT where 1) participants had neutral feelings concerning uneasiness, and 2) lateral control was not affected, was 7.9 seconds. Further findings reported by Asoh and Iihoshi (2006) suggested that TSOT had a stronger correlation with TGT (0.8933) than TTT (0.8760) for bench testing. The bench testing intervals for closed and open occlusion times in seconds were closed 0.5/opened 1.0, closed 1.0/opened 1.0, closed 0.5/opened 1.5, closed 1.0/opened 1.5, and closed 3.0/opened 1.5, with the closed 1.0/opened 1.5 alternative exhibiting the highest correlation. Based on these data, a TSOT of 7.5 seconds was found to be equivalent to 8 seconds TGT, which in turn was the TGT that resulted in neutral uneasiness levels.
Angell, Auflick, Austria, Kochhar, Tijerina, Biever, Diptiman, Hogsett, and Kiger (2006)	Conducted a review of various rules proposed for whether tasks were safe to complete while driving. They first interviewed experts in the field of driving and distraction as to whether certain task types were felt to need lower or higher levels of attention. They then ran a study having participants perform those tasks while in a static environment and recorded the time needed to complete tasks when vision was occluded and also when no occlusion goggles were used. Angell et al. (2006) compared the findings to what the experts agreed upon. The authors found that the best correspondence between different task classification rules and the expert's opinion was for a rule where a higher workload category was assigned to tasks whose TSOT exceeded 7.5 seconds. Angell et al. (2006) also showed correlations higher than 0.8 between TSOT and TGT.
Monk and Kidd (2007)	The authors conclude that the R value may be overestimated when participants do not have a tracking task to complete during the occlusion interval. It is also reported that "estimations of visual demand based on individual post-occlusion resumption times may provide a more precise measure of transition costs and resumability than measures based on Total Shutter Open Time." (p. 2). They propose an alternative formula for evaluating visual demand costs, using measurements of resumption lags and the times between unoccluded task inputs during each unoccluded episode. For their study, R and the proposed measures were not highly correlated. The authors state that this might be due to the small sample size of eight participants. They suggest that D may provide a more precise estimate of resumability because it is based on resumption measures.
Horberry, Stevens, Burnett, Cotter, and	State that guidelines should be developed for a detailed measurement protocol using standard occlusion. They specifically addressed the issue of age in relation to In-Vehicle Information System (IVIS) tasks. Their results showed differences in age with

Robbins (2008)	respect to TSOT and R. The older participants had more varied scores, especially for TSOT, and generally did more poorly than younger participants. The authors therefore suggest that to "obtain minimal inter-subject variability measurements [one] should use younger/middle-aged participants (broadly 17-66 years old)" (p. 175). The authors emphasize that this range merely reduces variability during testing and that the recommendation should not be taken as a suggestion that IVIS systems are developed only for younger drivers.
Gelau, Henning and Krems (2009)	Looked at how reliable the occlusion technique is for assessing visual distraction with IVIS. They looked at 4 different experiments that were previously conducted and looked at the internal consistency based on the TSOT. Their findings showed that "characteristic values for internal consistency were almost all in the range of .90 for the occlusion method and can be considered satisfactory."
Brook-Carter, Stevens, Reed, and Thompson (2009)	Looked at issues related to research gaps on system response delay, participant training levels prior to a study, number of repetitions for each task, sample size required, and calculating resumability. The authors suggest that three or more training tasks should be used for the more complex tasks being tested (e.g., place of interest search). Tasks that are less complex would require fewer training sessions. Their results also showed no significant difference with respect to the occluded versus non-occluded training method. Therefore, training with or without occlusion goggles shouldn't have a difference in the outcome of the research findings. In addition, based on differences between task completion times for successive trials of the same task, the authors recommend the use of 2 trials. In terms of sample size, the results suggested the use of over 10 participants for some tasks, although other tasks would benefit from a minimum sample size of 14 participants. The authors suggest a potential tradeoff of increasing the number of test trials rather than the sample size, which could be less time-consuming and expensive. Related to statistics, they suggest that "the value of R should be calculated on a within participant basis by taking the ratio of the respective median TSOT and TTTstatic across trial repetitions completed by one participant on each task under evaluation. Also, for the purpose of identifying interface designs that exhibit poor resumability, the 85th percentile value is suggested as the most appropriate"(p. 185). The use of the 85th percentile reduces the influence of outliers on the data.
Gelau and Schindhelm (2010)	Conducted a study having participants do a simple auditory tracking task during the occlusion period. Their findings supported those of Monk and Kidd (2007), discussed earlier. There was a significant increase in the TSOT and R value when participants performed the auditory tracking task. TSOT and R value measures obtained under an auditory tracking task also showed a clearer discrimination between an "easy" versus a "difficult" IVIS task than those obtained under non-auditory tracking task conditions. Furthermore, the R value was found to only be sensitive to the level of difficulty for the IVIS task when the auditory tracking task was used. The authors propose amending the standard occlusion method by having it include a cognitive load task, which they believe would improve the sensitivity of the method.
Stevens, Burnett, and Horberry (2010)	Point out that the current metric for the occlusion method lacks validation and does not "constrain measurement variability as much as it could" (p. 530). The authors believe the ISO standard is a good starting point but does not include all the measurement

	details. They proposed a new metric to test IVIS tasks that does not use the R value: Mean TSOT + (85th percentile – Mean)^2/Mean < 8.0s They justify the use of the 85th percentile because occlusion data are not usually normally distributed but instead follow a log-normal curve. Stevens et al. (2010) state that using the 85th percentile allows for more representation among participants who had more difficulty with some tasks. The use of a <8.9s TSOT is suggested as a conservative measure, and is based on JAMA's mean value of 7.5s. The authors adjusted that number to 8.0s to account for both the mean and the spread of the data.

G. In light of these findings, do you think these occlusion metrics are sufficient to describe visual demand for verification purposes? Why or why not?

H. Should the parameters for the occlusion test be adjusted and why?

   i. Could there be an influence from limiting the maximum glance duration at 1.5 seconds?

I. Is the 15 seconds TSOT excessive? Not sufficient? Why?

   i. Do you consider the 15 seconds TSOT threshold to be equivalent to the 2 / 20 seconds thresholds? Why or why not?

J. What would be appropriate (minimal) methods for statistical comparisons to provide reasonable statistical confidence and why?

K. Statistically speaking, what should be considered a sufficient sample size for these (or any other suggested) metrics in order to establish compliance?

   i. How many repetitions should be obtained per participant and why?

L. Do you consider occlusion as an early test, a final verification test, or somewhere between these two extremes? Can you provide reasons why you place it where you do?

   i. Do you think occlusion tests should be performed multiple times in the design cycle? How often?

VI. CLOSING QUESTIONS

Is there anything that you would like to add to the discussion that we've had today? That is, is there some topic or issue that we did not include in this questionnaire that you feel might be valuable?

Would you like to have your name included as a contributor to this research in any report or publication that may emanate from this work? (Please note that if we were to list your name, it would not be identified with any particular response or responses – rather, the list of contributors would be separated from the aggregated data and provided as part of an acknowledgement.)

RESPONSE: YES: ___ NO: ___

Thank you *so very much* for your time and your contributions!

DOT HS 811 763
May 2013

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

www.nhtsa.gov

8685-050613-v2