Report #16-UT-045

NSTSCE

National Surface Transportation Safety Center for Excellence

Comparing Three Methods of Eye-Glance Transition Coding

Julie A. McClafferty • Jon M. Hankey

Submitted: September 30, 2016

Technology
Impairment
Infrastructure
Mobility

Housed at the Virginia Tech Transportation Institute
3500 Transportation Research Plaza • Blacksburg, Virginia 24061

ACKNOWLEDGMENTS

The authors of this report would like to acknowledge the support of the stakeholders of the National Surface Transportation Safety Center for Excellence (NSTSCE): Tom Dingus from the Virginia Tech Transportation Institute, John Capp from General Motors Corporation, Chris Hayes from Travelers Insurance, Martin Walker from the Federal Motor Carrier Safety Administration, and Cathy McGhee from the Virginia Department of Transportation and the Virginia Transportation Research Council.

The NSTSCE stakeholders have jointly funded this research for the purpose of developing and disseminating advanced transportation safety techniques and innovations.

EXECUTIVE SUMMARY

In order for researchers to analyze driver glance behavior and compare their results to other studies, glance evaluation methodologies must be comparable in terms of how specific terms are defined. To this end, the International Organization for Standardization (ISO) developed the ISO 15007 standard (2014a, 2014b), which was adopted by the Society of Automotive Engineers (SAE) as SAE J2396 (2000). This report describes an experiment designed to test the validity and flexibility of one key aspect of these standards with respect to how transitions between glances (e.g., from the forward roadway to rearview mirror) may be assessed.

A glance dwell is the period of time that a driver's eyes are directed toward a specific direction or location (or area of interest) in a single glance, such as the forward roadway. Transitions are the time a driver spends moving from one area of interest to another (e.g., forward roadway and rearview mirror), or the time between dwells. Glance transitions (generally less than 0.1 to 0.5 seconds in duration each) make up a small percentage of the overall time in a given driver's glance pattern, but are a crucial aspect of glance behavior because they allow a driver to take in information from more than one area of interest.

The ISO 15007 and SAE J2396 standards define glance duration as the period of time that a driver's attention is on an area of interest, and specifies that transitions should be included with the destination location (the Destination Method, DM) to reflect the attentional shift toward the new glance location. According to the ISO standard, that attentional shift starts at the point where the leading transition begins. The DM requirement poses challenges to manual eye-glance coding in that it requires the analyst or analysis algorithm to either predict the glance destination in advance or to effectively code the glances in reverse (coding the transition after the destination is identified). In 2014, ISO published an updated standard that allows an option to include transitions with the origin glance duration (the Origin Method, OM) if the transitions to and from a destination can be assumed to be of equal length, but indicates that this is a less-preferred method. However, the OM appears to be a more efficient mode of coding and can provide considerable time and cost advantages in manual coding.

A third potential method is the Transition Method (TM), in which transitions are kept as separate glance entities and can be analyzed either as self-contained elements or as part of the destination or origin glances (making both DM and OM possible from the same data set). The TM also allows for calculation of additional metrics that are not possible with either DM or OM such as diversion duration, transition duration, and percentage of transition time. These additional metrics may be useful in some studies (e.g., as potential surrogates for driver fatigue, impairment, or cognitive load). The TM is also the only method capable of producing a clean dwell time metric.

Both TM and DM are likely to significantly increase both the error rate and the time required to manually code a given epoch compared to OM. This report does not delve into the underlying assumptions of the three methods (e.g., whether the driver's attention has shifted to the destination or remains with the origin during transitions), but concentrates instead on the practical ramifications of the three coding methods in terms of the time (and therefore cost) required, their effect on inter-rater reliability, and their effect on key glance metrics, including mean glance duration and total eyes off road time (TEORT). The goal of the project was to determine if there is a measurable difference in the output of the three methods when manual eye-glance coding is performed, and whether there are justifiable grounds to recommend or require one method over the others.

To test these three methods, 90 one-minute epochs were selected from the Second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study data set to include a variety of environmental, secondary task, and situational challenges to eye-glance analysis. These epochs were divided into three groups of 30 epochs each such that each group of epochs had a similar level of complexity. The three groups were then assigned to one of the three eye-glance methods. Five trained eye-glance analysts and one expert rater coded the three sets of epochs using the assigned method, and the resulting glance data were then analyzed to compare the time required to code for each method, the inter- and expert-rater reliability of the three methods, and the effect of coding method on key glance metrics.

Although the experiment reported here was small in terms of number of events and raters, the results are compelling and support the premise that the three methods for coding eye glances produce very similar results and reliability, but require very different degrees of resources. The key findings are listed below:

1. The OM is the fastest manual eye-glance coding method (with a recommended time multiplier of 10:1), followed by the DM (12:1), and then the TM (14:1). The amount of time it takes to code is directly related to cost.

2. All three methods produce similar inter-rater and expert-rater reliability scores, averaging about 88% before additional quality control measures are applied. Reliability may still be a factor if analyst experience with general eye-glance coding procedures is insufficient and/or previous coding accuracy is lower than desired. A quality control process should be a part of any coding effort, regardless of the selected method.

3. The three methods of coding yield similar key glance metrics, such as average percentage of time not forward, mean forward glance duration, and mean not forward glance duration.

In conclusion, OM is the most time and cost-efficient method for manual glance coding and produces key glance metrics that are comparable to those produced by the other, more costly methods. The added time and expense of the DM and TM without a corresponding increase in reliability or change in key metric values does not justify a requirement for their use over the OM if glance duration and metrics closely related to glance duration are the focus of the research question and planned analysis. One exception to this conclusion is that the TM represents the only method of the three that can provide the necessary data for some glance metrics specific to transitions (e.g., transition duration and frequency), and the added time and costs to perform the TM would be a necessary part of any research budget where transition-related metrics are essential.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

DM	Destination Method (including transition time with the destination glance)
ISO	International Organization for Standardization
OM	Origin Method (including transition time with the origin glance)
SAE	Society of Automotive Engineers
SHRP 2	Second Strategic Highway Research Program
TM	Transition Method (coding between-glance transitions as separate entities)
VTTI	Virginia Tech Transportation Institute

CHAPTER 1. INTRODUCTION AND JUSTIFICATION

Safe driving relies heavily on a driver's ability to receive and process visual information quickly and accurately in order to guide decision-making and vehicle maneuvering in an ever-changing and unpredictable driving environment. Many transportation researchers use assessments of glance location and scanning behavior to evaluate the quantity, content, and quality of the information being processed by individual drivers. The information processed by a driver is influenced by the number of glances and time spent glancing toward (1) the forward roadway, (2) other driving-related gaze directions (e.g., mirrors, speedometer), and (3) a wide variety of non-driving-related directions or secondary tasks (e.g., cell phone, passenger, roadside objects).

In order for each assessment to provide reliable conclusions that are comparable to other studies and that contribute to the broader research topic, glance evaluation methodology must be clear on several key properties. These key properties include the definitions of specific glance locations, the definitions of specific terms, and the method by which glance locations are determined (whether manual or automated, real-time or post hoc). To this end, the International Organization for Standardization (ISO) published the ISO 15007 standards in 2002 and revised them in (2014a, 2014b). An early committee draft of these standards was also adopted by the Society of Automotive Engineers (SAE) in SAE J2396 (2000).

This report describes an experiment designed to test the validity of one key aspect of these standards: how transitions between glances (e.g., from the forward roadway to rearview mirror) may be assessed.

DRIVER GLANCE TERMINOLOGY AND CODING METHODS

A <u>glance dwell</u> is the period of time that a driver's eyes are directed toward a specific direction or location (or <u>area of interest</u>) in a single glance, such as the forward roadway. <u>Transitions</u> are the time a driver spends moving from one area of interest to another (e.g., forward roadway and rearview mirror), or the time between dwells. Glance transitions (generally less than 0.1 to 0.5 seconds in duration each) make up an overall small percentage of time in a given driver's glance pattern, but are a crucial aspect of glance behavior as they allow a driver to take in information from more than one area of interest. <u>Glance duration</u> indicates the period of time that a driver's attention is on an area of interest, and may or may not include transition time, depending on the definition applied.

The ISO and SAE standards provide guidance on to how to incorporate glance transitions into key glance metrics, such as glance duration. However, despite the published standards, debate continues in the technical literature regarding the best methodology to handle these transitions (and even if there is a "best" way at all).

Until 2014, the ISO standard (adopted by the SAE) had specified that any time spent in transition between one glance location or area of interest (the origin) and another (the destination) was to be considered part of the destination glance duration. As stated in the standard, "this is done to reflect, or capture, the shift of attention toward the point of dwell. That attentional shift starts at the point where the leading transition begins." This method of including transitions with the destination glance is referred to in this report as the Destination Method (DM).

This DM requirement poses challenges in some methods of glance coding because it requires the analyst or analysis algorithm to either predict the glance destination in advance or to effectively code the glances in reverse (coding the transition after the destination is identified). The essence of the problem is that the destination glance cannot be known until after the driver has completed the transition and fixated there.

In 2014, ISO published an updated standard that still recommends the DM, but allows for the transition time to instead be considered as part of the origin glance duration "if the leading and trailing transitions are of equivalent length." This allows for an alternate calculation of glance duration in cases where the DM may not be feasible for a given assessment method. The standard goes on to say that "this is not ideal (given that it does not reflect the close relationship of attention and eyes) – but is an acceptable measurement compromise if applied carefully – one that should not produce any discernible difference in glance lengths (as long as leading and trailing transitions are of equivalent length)." Thus, the alternate method offered by the standard is regarded as an acceptable way to overcome some of the logistical challenges posed by the DM but is still considered inferior.

A review of the literature regarding the cognitive implications of transition time is beyond the scope of this report. However, the underlying assumptions about the relationship between attention and glance are still under debate. As stated above, the DM is recommended because it is believed that attention shifts to a new location once a person begins to transition away from a glance location, even though the person has not yet seen that location and has not yet received any visual information from that location. On the other hand, one could argue that the person is still processing the visual information received from the origin glance destination until new visual information is received from the destination location. Thus, it may be more appropriate to include the transition time with the origin glance to reflect visual information that has actually been received.

The method of including transitions in the origin glance duration is referred to in this report as the Origin Method (OM). This method has been posed as a potentially more efficient method of manually coding glance information due to the time series nature of video (though this assumption had never been tested before now), and the practical difference in the interpretation of the two methods is blurry.

Figure 1 illustrates the difference between the DM and OM methods for a scenario in which the driver shifts attention from the forward roadway to an object and back to the forward roadway. Glance duration for the DM is denoted by *c*, which consists of the transition time away from the forward roadway (*b*) and the dwell time on the object (*a*). Glance duration for the OM is denoted by *d*, which consists of the dwell time on the object (*a*) and the transition time from the object back to the roadway (*b*).

A third potential method is the Transition Method (TM), in which transitions are kept as separate glance entities and can be analyzed either as self-contained elements or as part of the destination or origin glances (making both DM and OM possible from the same data set). The TM also allows for calculation of additional metrics that are not possible with either DM or OM, such as diversion duration, transition duration, and the percentage of transition time. <u>Diversion duration</u> (illustrated in Figure 1 as *e*) is generally only calculated for non-forward glances and is the glance dwell time on a specified area of interest (such as a cell phone) plus the transition time both to and from that location. In essence, diversion duration represents the entire time that the eyes are not directed at a key location such as the forward driving scene. Transition duration (also mean transition duration) refers to the duration of each individual transition. Percentage transition time refers to the percentage of time spent in transition rather than dwelling on any one location over a given period of time. These additional metrics may be useful in some studies (e.g., as potential surrogates for driver fatigue, impairment, or cognitive load). The TM is also the only method capable of producing a clean dwell time metric.

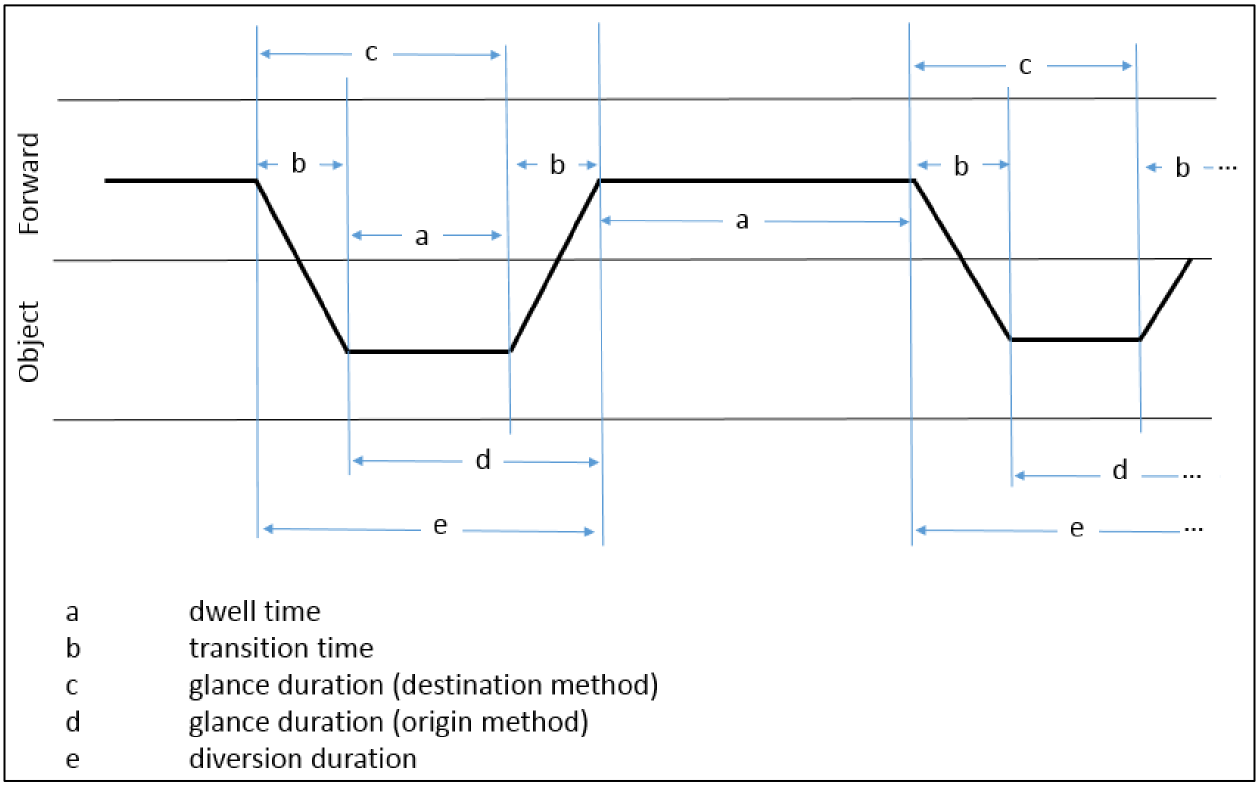


Figure 1. Illustration. Differences in the calculation of glance duration between DM and OM. TM is capable of producing both types of glance duration as well as diversion duration and transition time. (Modified from ISO 2014a)

LOGISTICS OF MANUAL CODING

The challenge of manual glance behavior coding (e.g., a human analyst manually codes the glances of another person without the aid of eye trackers) presents logistical and practical tradeoffs between these methods of transition analysis. Manual glance coding requires a trained analyst to view recorded video frame-by-frame and use context to determine the direction and/or object of each glance.

1. Following the OM, the analyst is able to move through the video in a mostly forward direction (i.e., time 1 to time 2 to time 3), stopping and reversing only to clarify specific frames where a new glance is fixated. This has generally been considered to

be the most efficient and cost-effective method, with a potentially lower error rate than other methods because of the increased efficiency.

2. Following the DM, the analyst must either code the video in reverse (i.e., time 3 to time 2 to time 1, so that the destination location is known before the transition is encountered) or code the video in the forward direction but then reverse the video every time a new destination glance is observed to code the previous transition. The DM is believed to be less efficient, more costly, and potentially more error-prone than the OM due to the non-intuitive reverse coding or frequent shifting between advancing and reversing video playback.

3. If the TM is used, transitions can be coded separately, and prior knowledge of the destination is not required. However, the TM also requires twice as many shifts in glance coding since both the start and the end of each transition must be identified. These two differences make it difficult to predict the effects of the TM on the rater time required compared to the other two methods. The increased burden of twice as many coding shifts also raises the concern of potentially increased error rates.

This report describes the results and methods of a small experiment designed to explore the costs and benefits of the three manual glance transition coding methods. Specifically, it looks at the time and cost required to perform each method, the effect of method selection on reliability and coding errors, and the impacts of method selection on key glance metrics, including mean glance duration and total eyes off road time (TEORT). This report does not delve into the underlying assumptions of the three methods (e.g., whether the driver's attention has shifted to the destination or remains with the origin during transitions), but is instead focused on determining if there is a measurable difference in the process and output of the three methods when manual eye-glance coding is performed, and whether there are justifiable grounds to recommend or require one method over the others.

CHAPTER 2. METHODS

DATA SOURCE

In order to compare the three methods of transition coding in naturalistic driver glance behavior analysis, specific epochs of data were selected from the Second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study (NDS) data set. The SHRP 2 data were collected between 2010 and 2013 from consented drivers located at six sites across the United States. Four video cameras were placed into each participant's vehicle to capture video of the driver's face, the forward roadway, the rear roadway, and the dashboard and driver's hands. Video was recorded at a frequency of 15 Hz (i.e., producing one frame every 66–67 ms). Glance information was obtained for this study primarily from the driver's face view, but the other views provided the necessary context to correctly identify glance locations. More information about this data set is available in the *Naturalistic Driving Study: Technical Coordination and Quality Control* report (Dingus et al., 2015).

EPOCH SELECTION

From the SHRP 2 NDS data set, 90 one-minute epochs were selected from consented driver trips to provide challenging coding scenarios and to represent the range of naturalistic conditions that would be expected in a larger sample. This conditional stratification provided for sufficient and above-average complexity in the test epochs and represented an above-average but equivalent coding challenge across all three coding methods. Since the test epochs were above average in complexity, it is believed that the results can be generalized to other less-complex naturalistic data sets.

Environmental conditions targeted in the set of test epochs included daytime and nighttime driving (different lighting conditions present unique coding challenges) as well as a cap on the number of interstate epochs (in order to ensure a variety of environmental complexities). Further, epochs were selected where the driver either did not wear glasses or did wear glasses (either lightly tinted or clear). Finally, in order to ensure that glances would be varied, epochs were selected to include a variety of secondary tasks such as cell phone use (and non-use), passenger interactions (and no interactions), center stack adjustments, and/or eating and drinking.

Once the 90 epochs were identified, they were split into three sets of 30 epochs each. Each set was assigned to one of the three coding methods. The three different event sets were needed in order to allow the same analysts to code events in all three methods (thereby controlling for rater experience) without coding the same event more than once (thereby preventing a familiarity bias in the time required). An alternative would have been to code the same 30 epochs with each method, but then three different sets of raters would be needed to avoid the familiarity bias, and having different raters could have resulted in other types of bias as well. Due to these tradeoffs, three separate (but similar) sets of events were selected, and the same group of raters coded each group using one of the three methods.

Similar distributions of the epoch selection criteria were represented in each epoch set. Table 1 below lists the characteristic breakdown of the three sets of 30 epochs. Note that the various selection criteria were not mutually exclusive; one epoch could meet multiple criteria (e.g., one epoch may meet the nighttime, passenger interaction, and non-interstate criteria), and all selected epochs met at least one criterion. In most cases, three epochs with similar characteristics (and the same driver) were sampled from the same trip and then assigned to the three different epoch sets. This further ensured that the three epoch sets would be similar in complexity.

Table 1. Characteristics of test epochs selected for the three transition coding methods. Thirty epochs were selected for each method, with a similar distribution of desired complexities.

Epoch Selection Criteria	# of Epochs (not mutually exclusive)		
	OM Set	**DM Set**	**TM Set**
Cell Phone Use	7	6	5
Passenger Interaction	13 (12)*	12 (11)	12 (11)
Center Stack Adjustment	7	4	5
Other Distraction	9	11	11
Lightly Tinted Glasses	6	6	6
Clear Glasses	17 (16)	16 (15)	16 (15)
No Glasses	6	7	7
Daytime	19 (18)	19 (18)	19 (18)
Nighttime	11	11	11
Non-Interstate Driving	23 (22)	23 (22)	23 (22)
Total # of epochs	30 (29)	30 (29)	30 (29)

*Numbers in parenthesis reflect the size of the subset after one epoch was removed from each set. Excessive sun glare prevented these epochs from producing usable glance data.

Prior to coding the test epochs, analysts were not familiar with the drivers included in the epoch set. However, due to the way the epochs were selected, each driver represented had one event in each of the three epoch sets. To avoid bias due to driver familiarity, the coding methods (and hence epoch sets) were assigned to raters in a random order. Each rater completed the sets in a different randomly assigned order, and the order of events within each set was also independently randomized.

After coding was complete on all three sets, one epoch was removed from each set (all three removed epochs were from the same driving trip). Excessive sun glare prevented the three epochs from that trip from producing usable glance data. Because epochs were coded in a randomized order by each analyst and this problematic file was not discovered until after coding was complete, the epoch could not be replaced and still maintain the randomized order. This resulted in a final test epoch set of 87 epochs, divided into three sets of 29.

ANALYST TRAINING AND GLANCE DEFINITIONS

Five trained eye-glance analysts plus one expert eye-glance analyst coded each of the three sets of epochs. The five analysts were considered an inter-rater test, and each was compared to the expert rater as well. These analysts each had at least 6 months of experience (the expert rater had over 12 months of experience) and a previously demonstrated exceptional accuracy (at least 90%

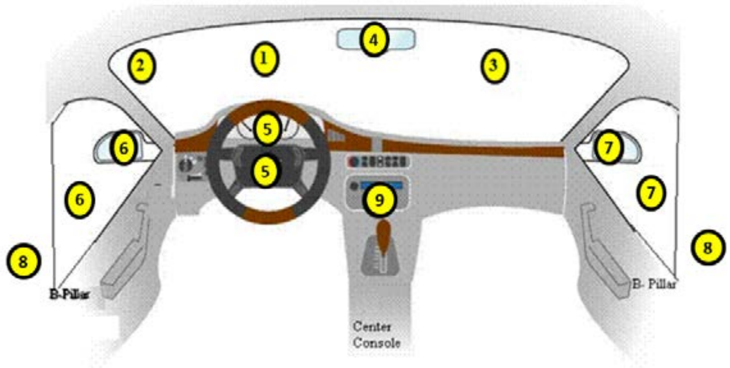
when compared to an expert rater for naturalistic data) in the OM of eye-glance coding for naturalistic data. (The OM was the standard method used at the Virginia Tech Transportation Institute [VTTI] at the time the test was conducted, and thus the method analysts were already familiar with). Each analyst was further trained in the DM and TM prior to this experiment, and each analyst was required to accumulate at least 8 hours of experience in these methods before starting on the test epochs. Data reductionists were limited to coding test epochs for no more than 4 hours per day, with mandatory 10-minute rest periods each hour to reduce fatigue-related errors.

Analysts coded the glances observed in each epoch to one of 18 categories, as listed in Table 2 and illustrated in Figure 2. Thirteen of these categories defined specific glance locations (e.g., Forward, Center Mirror, Passenger). The remaining five options could be coded during video frames when none of 13 discrete glance locations could be applied. These included “Eyes Closed” (to be used only when eyes were closed more than 5 consecutive frames or ~300 ms), “No Video” (to be used when two or more sequential frames of video were missing), two “Eyes Not Visible” options (to be used when eyes were not visible but were either known or not known to be off the forward roadway based on other context), and an “Other” option (for glances that could be identified but did not fall into any other category). Finally, a “Transition” category was included for the TM test cases only. Analysts used all the context available to them in the face, forward, hands, and rear video to determined glance locations for each frame of data within each test epoch. Appendix A provides complete definitions of the 18 glance location categories.

Minimum dwell time coded by analysts was 2 video frames. That is, the same glance location must be observed for at least 2 consecutive video frames in order to be coded as such. At the 15-Hz video frame rate available in the SHRP 2 NDS data set, this is equivalent to approximately 100–120 ms. Any glance shorter than this would be observed only as part of a transition and coded according to the transition protocol in use (OM, DM, or TM).

Table 2. List of glance locations coded.

Forward (Center)	Cell Phone
Left Forward	Portable Media Device
Right Forward	Interior Object
Rearview Mirror	Passenger
Left Window/Mirror	No Video
Right Window/Mirror	No Eyes Visible – Glance Location Unknown
Over-the-Shoulder (left or right)	No Eyes Visible – Eyes Are Off-Road
Instrument Cluster	Eyes Closed
Center Stack	Other
	*Transition (TM Only)



Key	Glance Location
1	Forward
2	Left Forward/Left Windshield
3	Right Forward/Right Windshield
4	Rearview Mirror
5	Instrument Cluster
6	Left Mirror/Left Window
7	Right Mirror/Right Window
8	Over-the-Shoulder (Left or Right)
9	Center Stack

Glance locations not shown: Transition, No Video, Passenger, Cell Phone, Portable Media Device, Interior Object, Other, No Eyes Visible – Glance Location Unknown, No Eyes Visible – Eyes Are Off-Road, Eyes Closed.

Figure 2. Diagram. Definitions of glance locations.

CODING INTERFACE

All epochs were coded using data reduction software developed by VTTI programmers to synchronize and display video and time-series driving data while providing an interface for data reduction. When coding glance data, analysts were able to play the video in a forward direction at various speeds, step forward and backward through video frames using arrow keys, and indicate glance locations at each video frame using assigned keyboard buttons. Coded data were then saved to the database for later extraction and analysis by the research team.

CHAPTER 3. RESULTS AND DISCUSSION

TIME AND COST COMPARISON

Figure 3 compares the average time required for analysts to complete the three methods of eye-glance coding per 1 minute of video. When coding using the OM, each minute of video required a mean of 11.0 minutes (SE = 1.5) of coding time across analysts. The DM and TM both required longer mean times than the OM (M = 12.5, SE = 1.6 and M = 14.2, SE = 2.4, respectively).

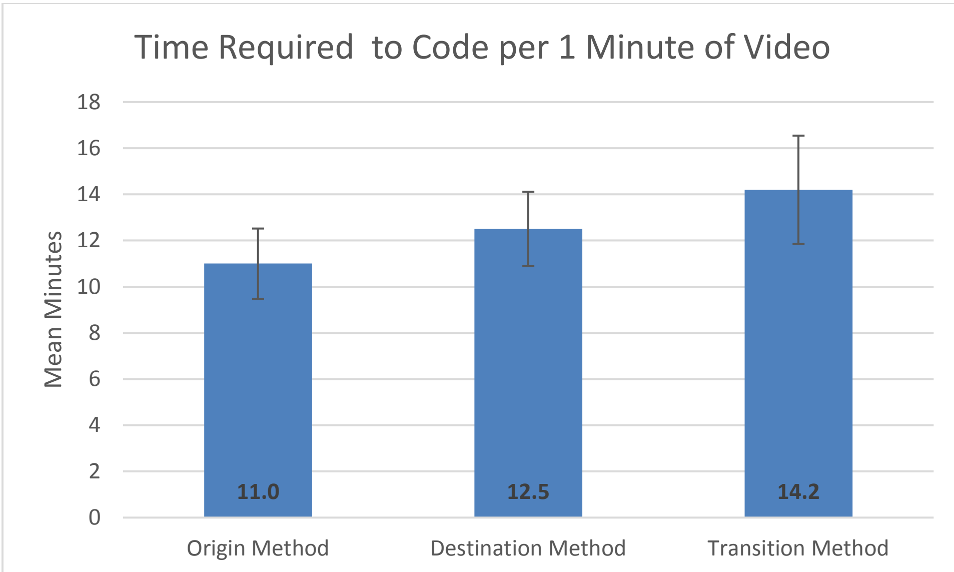


Figure 3. Chart. Comparison of time required to code the three methods of eye glance, in minutes of coding time per minute of video, averaged across raters.

Stated more simply, the DM coding required approximately 15% longer than the OM, and the TM required approximately 30% longer than the OM. The increased time for the DM reflects the "reverse coding" that is required for transitions in that method where the analyst moves forward in the video to determine the glance destination, and then reverses the video up to several frames to identify the start of the transition and code it. One solution to this dilemma would be to code fully in reverse (so that the destination is viewed before the transition is reached); however, a custom video reduction tool (which was not available) is required for this, and it may present additional glance interpretation challenges due to the non-intuitive nature of viewing video in reverse. The even longer time required for the TM reflects the need to pinpoint the video frames of twice as many coding shifts.

Regardless of the method selected, each shift in coding requires some degree of toggling between video frames in order to determine the exact time at which a new code should be applied (e.g., when the eyes begin moving to a new target or when the eyes fixate on a new target). This frame identification is one of the most time-consuming parts of any manual eye-glance coding process. In the DM and OM, only one shift in coding is required per glance transition (from origin to destination), with the differences between the two methods being where that shift is made (before or after the transition) and whether the destination has to be known prior to the transition being coded. However, in the TM, two coding shifts are required for every transition (origin to transition, transition to destination). As a result, despite being able to code the video in a mostly forward fashion (the destination is not required to be known before coding the transition), the time to code using the TM is longer than both the OM and DM.

Note that the standard errors all overlapped for the mean times across the three methods, indicating a lack of statistical significance. However, if the sample of epochs had been larger (magnifying the total time required) or the number of raters increased, it is likely that the significance would improve given the strong trend observed here.

One final observation can be taken from these data: the variation in the time required tends to be higher for the DM and TM (with TM having the highest standard deviation). This may have several implications. First, the increasing variance may be due to the lesser degree of experience the coders had with these two methods and the varying rates at which the analysts became acclimated to new methods. If this is the case, then the variation should decrease over time as analysts become more experienced.

An alternate explanation is that the DM and TM may have a greater variation in time required across analysts. If further research supports this idea, then accurately estimating the time and cost required to perform these types of eye-glance analyses (especially the TM) for the purpose of research planning and funding will be more difficult than it is for the OM.

The time to code eye-glance data is linearly and directly related to cost. Therefore, when preparing research proposals and/or designing research plans, the time to code should feed directly into the estimated costs. Based on the data presented here, and with the understanding that the epochs elected for this experiment were purposely more complex on average than what would normally be found in naturalistic data, it is recommended that the OM use an initial time multiplier per minute of video to be coded of 10:1, the DM use 12:1, and the TM use 14:1, assuming that the intended analysts have at least some background in eye-glance coding. These initial ratios allow researchers to estimate how long the initial coding will take for one rater on each epoch. Once the method and initial time multiplier are decided upon, additional time and costs must be budgeted to allow for the desired level of quality control. If dual coding is required, then these ratios should be multiplied by two. If a certain percentage will be "spot checked," then the ratios must be increased by that percentage.

A final note is necessary on the time and cost increases with the DM and TM. While rater inexperience may have played a role in the increased time for DM and TM methods observed in this experiment (recall that analysts were most experienced in the OM at the time of the experiment), the experience effect appears to be minimal when analysts are already well trained in the basic mechanics of glance coding. As evidence of this, as of July 2016 the TM has been the new VTTI standard eye-glance protocol for approximately 18 months, and the time required to code each minute of video using the TM has decreased only slightly below the rates observed in this experiment. The multiplier currently observed in VTTI's data reduction labs hovers around the 12:1 mark for the TM when applied to naturalistic data of average complexity. In contrast, prior to adopting the TM as a standard protocol, the OM had been VTTI's standard for more than 5 years, and the time multiplier remained consistently around 8:1 for average-complexity epochs. Note that these ratios are somewhat lower than the recommended budgeting ratios because they come from average-complexity events, whereas the events in the experiment presented here were intentionally more complex than average. In addition, these lower ratios are likely attainable on a consistent basis only by highly experienced glance analysts who perform these analyses regularly. The higher ratios recommended (10:1 for OM, 12:1 for DM, and 14:1 for TM) are more realistic for most research efforts.

RELIABILITY COMPARISON

Both inter-rater and expert-rater reliability were similar across the three coding methods (Table 3). The percentage agreement for the inter-rater reliability comparisons ranged from 87% to 90% for the OM (average 87%), 89% to 92% for the DM (average 90%), and 87% to 89% for the TM (average 88%). Similarly, expert-rater percentage agreement averaged 88% for all three methods. These scores strongly indicate that similar levels of reliability can be obtained with all the methods tested (provided that the required time is allowed).

Appendix B and Appendix C provide more detailed epoch-level expert-rater and inter-rater scores, respectively.

Table 3. Comparison summary of inter- and expert-rater reliability across three methods of eye-glance coding.

	% Agreement, Inter-Rater, Across Epochs						% Agreement, Expert Rater, Across Epochs
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average	
OM							
Rater 1	.	88%	89%	86%	87%	87%	86%
Rater 2	.	.	90%	87%	88%	89%	88%
Rater 3	.	.	.	89%	89%	89%	89%
Rater 4	.	.	.	.	89%	88%	87%
Rater 5	.	.	.	.	.	88%	87%
Overall Average	.	.	.	.	.	88%	87%
DM							
Rater 1	.	92%	89%	90%	91%	91%	88%
Rater 2	.	.	89%	90%	91%	90%	88%
Rater 3	.	.	.	90%	90%	90%	88%
Rater 4	.	.	.	.	91%	90%	89%
Rater 5	.	.	.	.	.	91%	89%
Overall Average	.	.	.	.	.	90%	88%
TM							
Rater 1	.	88%	89%	89%	87%	88%	88%
Rater 2	.	.	88%	87%	88%	88%	88%
Rater 3	.	.	.	88%	88%	88%	88%
Rater 4	.	.	.	.	88%	88%	88%
Rater 5	.	.	.	.	.	88%	88%
Overall Average	.	.	.	.	.	88%	88%

Under VTTI data reduction standards, reduced glance data would normally undergo a quality control process whereby a second experienced person with demonstrated high reliability would review some or all of the coded data prior to delivery to the research team for statistical analysis. Disagreements would then be resolved between the original analyst and the second reviewer, and a senior analyst would be included if needed. The glance data coded during this experience did not go through this iterative process in order to preserve the true inter- and expert-rater scores. As a result, the accuracy of the glance coding reported here is likely lower than it would otherwise be across all three methods.

COMPARISON OF KEY METRICS

Once the time and cost to complete and reliability analysis were completed, the final step in the eye-glance methodology assessment was to examine the impact of the selected methodology on several key glance metrics.

Transition Duration and Percentage of Time in Transition for TM

To set the stage for examining other key glance metrics, summary statistics were first calculated on the data generated by the TM regarding mean transition duration and the percentage of time spent in transition. (Recall that these metrics can only be calculated if the TM method is followed.) Table 4 summarizes the transition data recorded by the expert rater across the 29 epochs assigned to the TM. Overall, transitions made up 5.6% (SE = 0.5%) of total epoch time on average, and the average transition duration across all transitions was 0.2 seconds (SE = 0.01). All transitions include those to and from Forward as well as those between two consecutive Non-Forward glances.

Table 4. Percentage of time in transition and mean transition duration for TM epochs.

	Across TM Epochs		Across TM Transitions	
	% Time Spent in Transition		Transition Duration (s)	
Transition Type	**Mean**	**SE**	**Mean**	**SE**
All Observed Transitions*	5.6%	0.5%	0.20	<0.01
Forward to Non-Forward Transitions	2.4%	0.2%	0.19	0.01
Non-Forward to Forward Transitions	2.8%	0.3%	0.19	<0.01

* All Observed Transitions includes transitions to and from Forward as well as between two Non-Forward glance locations.

One stipulation of the ISO standard states that transitions may be added to the origin location to calculate glance duration only if the leading and trailing transitions are (or can be assumed to be) of equivalent length. To confirm this, both mean transition duration and the percentage of time in transition were broken down by whether each transition was away from Forward (eyes moving from Forward to Non-Forward) or toward Forward (eyes moving from Non-Forward to Forward). The mean transition duration for these two transition types was 0.19 seconds (SE = 0.01) and 0.19 seconds (SE < 0.01) respectively, and the percentage of total time spent in one of these two types of transitions was also similar at 2.4% (SE = 0.2%) for "from Forward" and 2.8% (SE = 0.3%) for "to Forward." It therefore seems reasonable to conclude that the leading and trailing transitions are equivalent in duration, providing initial support for equal acceptance of including transition time with either the origin or the destination.

It is important to recognize that not all glance transitions result in transition data. If a transition occurs so quickly that the eyes appear fixated on one location in one video frame and on another location in the next video frame, no transition frames will be coded, and thus there will be no transition data to include in these metrics. In these cases, the driver's eyes transition between video frames, resulting in an unknown transition duration that falls somewhere between 0 ms and 67 ms (assuming 15-Hz video). One strategy for handling unobservable transitions is to select an arbitrary surrogate duration (e.g., 0 ms, 67 ms, or somewhere in-between), which would have varying effects on the resulting summary measures. For example, if these transitions are considered to have a duration of 0 ms, then mean transition durations would be disproportionately lowered (and underestimated) while the percentage of time in transition would remain unchanged. Other surrogate values could result in either under- or overestimating both mean duration and total percentage of time. The problem is that it would be impossible to know whether the result was an under- or overestimate.

An alternate, more conservative strategy for handling unobservable transitions, and the strategy used here, is to exclude them from mean duration and total percentage of time calculations. The result is that the mean transition duration is known to be a slight overestimation, and the total percentage of time in transition is known to be a slight underestimation.

Comparison of Glance Duration Metrics

After three sets of epochs had been coded according to their assigned method (OM, DM, or TM), several key glance metrics were calculated for comparison between methods. These results are listed in Table 5.

Table 5. Comparison of key glance metrics between three eye-glance coding methods. Each method was applied to one of three sets of 29 one-minute epochs with similar environmental and behavioral characteristics.

		Event Set	% Time Eyes Not Forward		Mean Forward Glance Duration (s)		Mean Duration Eyes Off Forward (s)	
			Mean	SE	Mean	SE	Mean	SE
	OM	**1**	26%	2.6%	2.6	0.4	1.0	0.06
	DM	**2**	24%	2.2%	3.1	0.2	1.0	0.03
TM	**Excluding Transitions**	***3***	16%	1.7%	3.9	0.3	0.72	0.02
	Transitions w/ Origin	***3***	18%	1.8%	4.2	0.4	0.93	0.03
	Transitions w/ Destination	***3***	18%	1.7%	4.0	0.3	0.92	0.03
	OM*	***3***	17%	1.7%	4.2	0.4	0.94	0.06
	DM*	***3***	18%	1.9%	3.9	0.4	0.88	0.06

* Event Set 3 was later coded with the OM and DM by two different sets of analysts to determine if differences were due to the method used or epochs selected.

Although the means of the metrics appear to vary somewhat for the three coding methods (the TM method in particular), the differences can largely be attributed to the different sets of epochs. While the levels of environmental and situational complexity and secondary task engagement by drivers were made as similar as possible between the three sets of epochs, it was impossible to make them identical and still maintain the independent nature of the three methods. The epochs were sampled from naturalistic data, which inevitably will include some uncontrolled variability.

Three aspects of the epoch data sets and resulting metrics lead to the conclusion that the methods produce similar, comparable results:

1. The OM and DM epochs resulted in very similar key glance metrics. When just these two epoch sets are examined, the mean percentages of time off road were 26% (SE = 2.6%) and 24% (SE = 2.2%), respectively, and mean forward glance durations were 2.6 s (SE = 0.4) and 3.1 s (SE = 0.2), respectively. Mean "off forward" durations were both 1.0 s (SE = 0.06 and 0.03, respectively).

2. Although the TM epochs produce mean values that appear to differ from the other methods, when the transitions from the TM epochs were analyzed in two different ways to mimic the OM and DM coding methods for the same set of epochs, the resulting metrics were nearly identical to each other. The average percentage of time not forward was 18% for both methods (SE = 1.8% and 1.7%, respectively), the average forward glance durations were 4.2 s (SE = 0.4) and 4.0 s (SE = 0.3), respectively, and the average non-forward durations were 0.93 s (SE = 0.03) and 0.92 s (SE = 0.03), respectively. The most logical conclusion is that the epochs assigned to the DM and OM methods were more similar to each other in terms of driver glance behavior than they were to the TM epochs, which highlights a sampling problem inherent in naturalistic data rather than a methodological problem inherent in any one of the coding methods.

   This conclusion was further tested by having a separate set of raters code Event Set 3 (the events assigned to the TM) using the OM and yet a different set of raters code Event Set 3 using the DM. Neither set of raters had coded these epochs previously. This produced essentially identical results to the TM analyzed either with transitions as part of the origin or destination glances. When coded using the OM or DM respectively, the average percentage of time not forward was 17% (SE = 1.7%) and 18% (SE = 1.9%), average forward glance duration was 4.2 s (SE = 0.4) and 3.9 s (SE = 0.4), and average non-forward duration was 0.94 s (SE = 0.06) and 0.88 s (SE = 0.06).

3. Even across the different and imperfectly sampled sets of epochs, the standard errors for each metric prevented these differences from being meaningful. No single mean or standard error combination in Table 5 stands out as creating methodological bias.

CHAPTER 4. CONCLUSIONS

Although the experiment reported here was small in terms of number of events and raters, the results are compelling and support the premise that the three methods for coding eye glances produce very similar results and reliability, but require very different degrees of resources. The key findings are listed below:

1. The OM is the fastest manual eye-glance coding method, followed by the DM and then the TM. Time to code is directly related to cost.

2. All three methods produce similar inter-rater and expert-rater reliability scores. Reliability may still be a factor if analyst experience with general eye-glance coding procedures is insufficient and/or previous coding accuracy is lower than desired. A quality control process should be a part of any coding effort, regardless of the selected method.

3. The three methods of coding will yield similar key glance metrics such as average percentage of time not forward, mean forward glance duration, and mean not forward glance duration.

The OM is the most time- and cost-efficient method for manual glance coding and produces key glance metrics that are comparable to those produced by the other, more costly methods. The added time and expense of the DM and TM without a corresponding increase in reliability or change in key metric values does not justify a requirement for their use over the OM if glance duration and metrics closely related to glance duration are the focus of the research question and planned analysis. One exception to this conclusion is that the TM represents the only method of the three that can provide the necessary data for some glance metrics specific to transitions (e.g., transition duration and frequency), and the added time and costs to perform the TM would be a necessary part of any research budget where transition-related metrics are essential.

APPENDIX A. GLANCE LOCATION DEFINITIONS

F	Forward (Center)	Any glance out the forward windshield directed towards the direction of the vehicle's travel. Note that when the vehicle is turning, these glances may not be directed directly forward but towards the vehicle's heading. Count these as forward glances.
D	Left Forward	Any glance out the forward windshield where the driver appears to be looking specifically out the left margin of the windshield (e.g., as if scanning for traffic before turning or glancing at oncoming traffic). This glance location includes any time the driver is looking out the windshield, but clearly not in the direction of travel (e.g., at road signs or buildings).
G	Right Forward	Any glance out the forward windshield where the driver appears to be looking specifically out the right side of the windshield (e.g., as if scanning for traffic before turning, at a vehicle ahead in an adjacent lane, or reading a road sign). This is often preceded or followed by Left Forward. This glance location includes any time the driver is looking out the windshield, but clearly not in the direction of travel (e.g., at road signs or buildings).
M	Rearview Mirror	Any glance to the rearview mirror or equipment located around it. This glance generally involves movement of the eyes to the right and up to the mirror. This includes glances that may be made to the rearview mirror in order to look at or interact with back-seat passengers.
L	Left Window/Mirror	Any glance to the left side mirror or window.
R	Right Window/Mirror	Any glance to the right side mirror or window.
S	Over-the-Shoulder (left or right)	Any glance over either of the participant's shoulders. In general, this will require the eyes to pass the B-pillar. If over the left shoulder, the eyes may not be visible, but this glance location can be inferred from context. NOTE: If it is clear from context that an over-the-shoulder glance is being made NOT to check a blind spot but instead to interact with a rear-seat passenger (e.g., food/toy is being handed back), then code the glance as Passenger. If context cannot be known with a high level of certainty, then code as Over-the-Shoulder.

I	Instrument Cluster	Any glance to the instrument cluster underneath the dashboard. This includes glances to the speedometer, control stalks, and steering wheel.
C	Center Stack	Any glance to the vehicle's center stack (vertical). Not to be confused with center console (cup holder area between driver and passenger), which is discussed under "Interior Object."
P	Cell Phone (electronic communications device)	Any glance at a cell phone or other electronic communications device (e.g., Blackberry), no matter where it is located. This includes glances to cell-phone-related equipment (e.g., battery chargers).
H	Portable Media Device	Any glance at a portable media device (e.g., MP3 player) or other personal digital music device, no matter where it is located.
W	Interior Object	Any glance to an identifiable object in the vehicle other than a cell phone. These objects include personal items brought in by the participant (e.g., purse, food, papers), any part of the participant's body (e.g., hand, ends of hair), electronic devices other than cell phones (e.g., laptop, PDA), and also OEM-installed devices that do not fall into other categories (e.g., door lock, window and seat controls). Glances to the center console (cup holder area between passenger seat and driver seat) will also be included in this category. The object does not need to be in the camera view for a specific frame to be coded with this category. If it is clear from surrounding video that the participant is looking at the object, this category may be used. This category can be used regardless of whether the participant's hands are or are not visible. NOTE: If the driver is looking at something that thepassenger is handing to them, code the eye glance as Passenger, until the object is fully in the driver's hand, then code as Interior Object (unless it is a cell phone, in which case code as Cell Phone). Also, if the driver is looking at something that the passenger is holding (but never hands to the driver), code as Passenger (not interior object). Individual studies may ask reductionists to identify objects in logs or drop-down menus, or may categorize specific objects as Systems of Interest.

A	Passenger	Any glance to a passenger, whether in front seat or rear seat of vehicle. You will need to use context (e.g., they're talking, or handing something) in order to determine this in some situations. NOTE: This does NOT include glances made to rear-seat passenger via the rearview mirror. Such glances should be coded as Rearview Mirror. NOTE: If the driver is looking at something that the passenger is handing to the driver, code the eye glance as Passenger, until the object is fully in the driver's hand, then code as Interior Object (unless it is a cell phone, in which case code as Cell Phone). Also, if the driver is looking at something that the passenger is holding (but never hands to the driver), code as Passenger glance (not Interior Object).
V	No Video	Unable to complete glance analysis because the face video view is temporarily unavailable. NOTE: This sometimes occurs for 1 or 2 frames at a time, and a "video not available" message may appear. If the glance location is the same before and after this occurs and the period is only 1 or 2 frames long, then code through this period as the glance location present before and after. If the "video not available" period is longer than 2 frames OR it occurs during a transition, use the "No Video" option.
E	No Eyes Visible – Glance Location Unknown	Unable to complete glance analysis due to an inability to see the driver's eyes and face. Video data are present, but the driver's eyes and face are not visible due to an obstruction (e.g., visor, hand,), or due to glare. Use this category when there is no way to tell whether the participant's eyes are on or off the road. This is the default and most often used "unknown" option, but there may be times with the "off road" option listed below may be appropriate.
T	No Eyes Visible – Eyes Are Off-Road	Unable to enter in specific glance location due to an inability to see the driver's eyes and face. However, it is clear that the participant is not looking at the roadway. Video is present, but the driver's eyes and face are not visible due to an obstruction (e.g., visor, hand), head position, or due to glare.

		Use this category when the eyes are not visible, you are not sure what the participant is looking at, but it is obvious that the eyes are not on the roadway.
Z	Eyes Closed	Any time that BOTH the participant's eyes are closed outside of normal blinking (e.g., the subject is falling asleep or rubbing eyes). As a rule of thumb, if the eyes are closed for five or more timestamps (1/3 of a second) during a slow blink, code it as Eyes Closed. Otherwise, code it as the glance location present before the eyes closed. If one eye remains open, code the location according to the open eye. If only one eye is visible, code according to the visible eye.
O	Other	Any glance that cannot be categorized using the above codes. If you come across anything that could fall under this category, please inform the Lab Manager for appropriate follow-up. Some pre-approved uses of the "other" option are listed below: • When the driver is looking forward, and then looks straight up at the sky as if watching a plane fly by. • When the driver is tilting head back to drink and the eyes leave the forward glance but do not really focus on anything at all.
Q	Transition	Any glance that is after the eyes stop focusing on an object (or forward) and before the re-affix on a new glance location. *This glance code is only valid for the Transition Method.*

APPENDIX B. EPOCH-LEVEL EXPERT-RATER SCORES

Table B1. Expert rater scores for the Origin Method across 29 epochs, 5 raters, and 1 expert rater.

Epoch Set 1	% Expert Rater Agreement for Each Rater					
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average for Epoch
1	66%	96%	--	76%	67%	76%
2	84%	92%	93%	92%	--	90%
3	82%	77%	83%	84%	86%	82%
4	90%	91%	88%	90%	87%	89%
5	94%	95%	94%	93%	92%	93%
6	88%	94%	95%	94%	93%	93%
7	--	93%	95%	93%	96%	94%
8	83%	81%	--	85%	82%	83%
9	88%	88%	93%	89%	88%	89%
10	99%	99%	98%	98%	97%	98%
11	78%	75%	76%	73%	73%	75%
12	92%	92%	--	89%	92%	91%
13	--	93%	92%	89%	86%	90%
14	--	80%	84%	83%	82%	82%
15	83%	86%	--	84%	79%	83%
16	95%	93%	95%	88%	87%	92%
17	94%	93%	--	95%	88%	93%
18	83%	83%	82%	79%	81%	82%
19	82%	87%	86%	84%	84%	85%
20	92%	90%	95%	86%	89%	90%
21	93%	95%	94%	89%	94%	93%
22	80%	84%	--	79%	86%	82%
23	--	95%	97%	95%	95%	95%
24	75%	73%	75%	72%	81%	75%
25	84%	83%	84%	82%	88%	84%
26	89%	87%	85%	89%	87%	88%
27	86%	90%	86%	86%	91%	88%
28	95%	95%	93%	94%	95%	94%
29	84%	87%	--	84%	86%	85%
Average for Rater	86%	88%	89%	87%	87%	87%
SE.	*1.5%*	*1.2%*	*1.4%*	*1.2%*	*1.3%*	*1.1%*

Table B2. Expert rater scores for the Destination Method across 29 epochs, 5 raters, and 1 expert rater.

Epoch Set 2	% Expert Rater Agreement for Each Rater					
	Rater 1	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Average for Epoch**
31	83%	87%	81%	84%	82%	83%
32	95%	95%	95%	97%	96%	95%
33	90%	90%	89%	87%	83%	88%
34	91%	91%	90%	91%	91%	91%
35	88%	91%	92%	93%	95%	92%
36	95%	95%	95%	96%	97%	96%
37	82%	83%	80%	83%	81%	81%
38	75%	69%	68%	75%	70%	72%
39	94%	95%	94%	95%	96%	95%
40	89%	90%	91%	90%	88%	90%
41	88%	87%	92%	87%	89%	88%
42	90%	85%	--	76%	82%	83%
43	93%	93%	85%	93%	95%	92%
44	97%	98%	98%	98%	99%	98%
45	82%	82%	--	79%	83%	81%
46	94%	91%	90%	92%	94%	92%
47	96%	95%	95%	96%	96%	96%
48	92%	92%	92%	91%	92%	92%
49	85%	86%	84%	85%	87%	86%
50	91%	89%	91%	92%	92%	91%
51	91%	88%	92%	91%	92%	91%
52	75%	78%	--	78%	82%	78%
53	75%	74%	73%	88%	91%	80%
54	66%	67%	73%	70%	69%	69%
55	83%	83%	90%	91%	95%	88%
56	85%	84%	83%	87%	88%	85%
57	93%	95%	92%	93%	95%	94%
58	93%	93%	91%	95%	96%	94%
59	95%	94%	--	98%	93%	95%
Average for Rater	88%	88%	88%	89%	89%	88%
SE	1.4%	1.5%	1.5%	1.4%	1.4%	1.3%

Table B3. Expert rater scores for the Transition Method across 29 epochs, 5 raters, and 1 expert rater.

Epoch Set 3	% Expert Rater Agreement for Each Rater					
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average for Epoch
61	86%	87%	84%	86%	85%	86%
62	94%	90%	91%	94%	94%	92%
63	81%	82%	83%	81%	80%	82%
64	90%	88%	92%	94%	89%	91%
65	96%	97%	95%	97%	91%	95%
66	95%	95%	95%	96%	95%	95%
67	80%	82%	85%	79%	87%	83%
68	78%	67%	65%	76%	70%	71%
69	80%	88%	88%	88%	--	86%
70	85%	77%	83%	78%	84%	81%
71	87%	88%	86%	90%	79%	86%
72	88%	90%	92%	91%	88%	90%
73	89%	86%	89%	82%	88%	87%
74	87%	90%	85%	91%	93%	89%
75	90%	87%	87%	91%	86%	88%
76	86%	88%	85%	88%	87%	87%
77	95%	93%	96%	93%	93%	94%
78	83%	87%	86%	84%	85%	85%
79	79%	85%	83%	83%	84%	83%
80	93%	90%	95%	88%	89%	91%
81	95%	95%	92%	93%	93%	94%
82	89%	90%	92%	86%	86%	88%
83	90%	96%	87%	94%	92%	92%
84	77%	70%	78%	74%	76%	75%
85	94%	95%	96%	96%	96%	95%
86	90%	91%	90%	85%	88%	89%
87	90%	92%	92%	93%	90%	92%
88	86%	84%	87%	88%	89%	87%
89	91%	95%	93%	94%	94%	93%
Average for Rater	88%	88%	88%	88%	88%	88%
SE	*1.0%*	*1.1%*	*1.2%*	*1.2%*	*1.3%*	*1.1%*

APPENDIX C. APPENDIX C. EPOCH-LEVEL INTER-RATER SCORES

Table C1. Inter-rater scores for the <u>Origin Method</u> across 29 epochs and 5 raters.

Epoch Set 1	% Inter-Rater Agreement for Indicated Raters										Average for Epoch
	1 vs. 2	1 vs. 3	1 vs. 4	1 vs. 5	2 vs. 3	2 vs. 4	2 vs. 5	3 vs. 4	3 vs. 5	4 vs. 5	
1	65%	--	67%	65%	--	75%	69%	--	--	73%	69%
2	79%	91%	91%	--	91%	93%	--	95%	--	--	90%
3	85%	88%	84%	82%	86%	83%	81%	85%	81%	76%	83%
4	91%	92%	91%	90%	90%	89%	91%	90%	91%	88%	90%
5	95%	97%	94%	96%	95%	94%	93%	95%	94%	92%	94%
6	89%	89%	90%	87%	93%	91%	95%	94%	93%	90%	91%
7	--	--	--	--	93%	93%	97%	89%	95%	91%	93%
8	83%	--	85%	83%	--	80%	82%	--	--	82%	82%
9	88%	88%	86%	90%	87%	89%	89%	91%	88%	87%	88%
10	100%	99%	98%	98%	99%	98%	98%	97%	98%	97%	98%
11	87%	87%	76%	81%	84%	79%	81%	78%	82%	71%	80%
12	99%	--	89%	94%	--	88%	93%	--	--	90%	92%
13	--	--	--	--	93%	87%	86%	87%	84%	86%	87%
14	--	--	--	--	92%	95%	89%	96%	90%	90%	92%
15	95%	--	92%	92%	--	96%	91%	--	--	88%	92%
16	93%	95%	88%	86%	93%	84%	91%	87%	86%	85%	89%
17	95%	--	95%	88%	--	95%	89%	--	--	88%	92%
18	93%	87%	90%	93%	88%	88%	90%	87%	87%	88%	89%
19	81%	88%	82%	89%	85%	88%	84%	85%	90%	84%	85%
20	92%	95%	85%	90%	91%	84%	87%	86%	92%	80%	88%
21	93%	94%	87%	93%	92%	87%	94%	90%	92%	89%	91%
22	83%	--	78%	78%	--	80%	81%	--	--	75%	79%
23	--	--	--	--	95%	96%	95%	95%	94%	94%	95%
24	67%	73%	80%	75%	69%	68%	72%	72%	74%	74%	72%
25	83%	78%	72%	79%	85%	78%	90%	80%	85%	81%	81%
26	95%	92%	94%	95%	92%	90%	96%	89%	91%	91%	92%
27	86%	83%	79%	84%	85%	83%	86%	85%	83%	84%	84%
28	95%	94%	94%	94%	95%	95%	96%	98%	94%	95%	95%
29	85%	--	74%	80%	--	81%	87%	--	--	83%	82%
Average for Pair	88%	89%	86%	87%	90%	87%	88%	89%	89%	85%	88%
SE	*1.7%*	*1.5%*	*1.6%*	*1.6%*	*1.3%*	*1.4%*	*1.3%*	*1.4%*	*1.2%*	*1.3%*	*1.3%*

Table C2. Inter-rater scores for the Destination Method across 29 epochs and 5 raters.

Epoch Set 2	% Inter-Rater Agreement for Indicated Raters										Average for Epoch
	1 vs. 2	1 vs. 3	1 vs. 4	1 vs. 5	2 vs. 3	2 vs. 4	2 vs. 5	3 vs. 4	3 vs. 5	4 vs. 5	
31	90%	84%	81%	87%	86%	81%	88%	78%	81%	79%	84%
32	97%	96%	97%	96%	96%	97%	96%	97%	95%	97%	96%
33	86%	87%	86%	79%	90%	90%	83%	89%	80%	85%	86%
34	98%	96%	96%	96%	96%	97%	96%	94%	94%	96%	96%
35	92%	90%	90%	91%	90%	93%	94%	92%	94%	95%	92%
36	95%	96%	96%	96%	96%	96%	97%	97%	97%	98%	97%
37	91%	91%	95%	91%	90%	93%	92%	94%	91%	93%	92%
38	80%	70%	73%	80%	64%	65%	75%	72%	67%	65%	71%
39	98%	98%	98%	98%	99%	99%	98%	98%	98%	99%	98%
40	98%	97%	98%	97%	99%	100%	96%	99%	96%	96%	98%
41	91%	92%	89%	90%	91%	94%	89%	90%	88%	90%	90%
42	86%	--	77%	87%	--	83%	83%	--	--	80%	83%
43	96%	85%	93%	92%	84%	92%	92%	89%	89%	96%	91%
44	98%	95%	96%	97%	96%	97%	98%	96%	97%	98%	97%
45	93%	--	89%	86%	--	93%	89%	--	--	86%	89%
46	93%	93%	96%	95%	94%	90%	90%	91%	89%	92%	92%
47	97%	94%	96%	98%	96%	97%	98%	96%	96%	98%	97%
48	98%	99%	95%	98%	100%	96%	99%	96%	99%	95%	97%
49	95%	88%	87%	91%	90%	86%	93%	85%	92%	87%	89%
50	93%	94%	98%	98%	94%	94%	92%	95%	95%	98%	95%
51	89%	90%	94%	94%	91%	89%	90%	89%	92%	95%	91%
52	90%	--	91%	88%	--	91%	86%	--	--	91%	90%
53	87%	58%	74%	79%	56%	70%	77%	75%	72%	84%	73%
54	71%	75%	73%	77%	72%	71%	76%	78%	77%	76%	75%
55	92%	85%	91%	82%	85%	91%	83%	86%	94%	89%	88%
56	92%	90%	93%	89%	92%	89%	92%	88%	90%	90%	91%
57	96%	95%	93%	98%	94%	95%	97%	91%	96%	93%	95%
58	90%	91%	94%	94%	90%	94%	95%	92%	92%	97%	93%
59	95%	--	94%	96%	--	92%	96%	--	--	92%	94%
Average for Pair	92%	89%	90%	91%	89%	90%	91%	90%	90%	91%	90%
SE	*1.1%*	*1.9%*	*1.4%*	*1.2%*	*2.1%*	*1.6%*	*1.3%*	*1.5%*	*1.7%*	*1.5%*	*1.4%*

Table C3. Inter-rater scores for the Transition Method across 29 epochs and 5 raters.

Epoch Set 3	% Inter-Rater Agreement for Indicated Raters										Average for Epoch
	1 vs. 2	**1 vs. 3**	**1 vs. 4**	**1 vs. 5**	**2 vs. 3**	**2 vs. 4**	**2 vs. 5**	**3 vs. 4**	**3 vs. 5**	**4 vs. 5**	
61	84%	83%	85%	83%	88%	84%	87%	84%	85%	86%	85%
62	92%	90%	93%	93%	87%	89%	90%	92%	92%	94%	91%
63	84%	87%	86%	81%	83%	80%	81%	85%	87%	82%	84%
64	84%	89%	92%	84%	87%	86%	90%	93%	88%	88%	88%
65	97%	95%	96%	90%	96%	97%	92%	96%	94%	88%	94%
66	95%	98%	96%	95%	94%	94%	95%	96%	94%	94%	95%
67	90%	85%	84%	87%	86%	83%	90%	83%	91%	87%	87%
68	68%	61%	79%	78%	65%	65%	64%	68%	64%	74%	69%
69	82%	83%	84%	--	94%	95%	--	97%	--	--	89%
70	70%	81%	76%	79%	71%	78%	74%	75%	80%	78%	76%
71	88%	90%	90%	80%	88%	88%	87%	90%	80%	79%	86%
72	94%	92%	90%	84%	92%	90%	86%	94%	88%	87%	90%
73	90%	92%	88%	89%	95%	90%	94%	90%	96%	89%	91%
74	87%	85%	90%	90%	86%	87%	91%	85%	84%	90%	87%
75	86%	90%	94%	86%	93%	85%	94%	89%	91%	84%	89%
76	93%	87%	92%	84%	88%	94%	86%	86%	81%	86%	88%
77	94%	98%	94%	93%	94%	92%	94%	94%	93%	93%	94%
78	85%	86%	80%	84%	87%	81%	85%	81%	85%	91%	85%
79	88%	90%	91%	87%	92%	89%	91%	91%	92%	90%	90%
80	94%	98%	92%	93%	94%	95%	96%	92%	94%	93%	94%
81	99%	93%	98%	96%	93%	97%	96%	94%	93%	94%	95%
82	97%	92%	83%	95%	91%	82%	93%	85%	88%	84%	89%
83	88%	91%	90%	91%	85%	91%	91%	87%	89%	91%	89%
84	71%	85%	77%	77%	74%	75%	74%	78%	80%	76%	77%
85	92%	94%	93%	93%	94%	94%	94%	96%	95%	95%	94%
86	90%	92%	90%	90%	91%	87%	90%	88%	93%	88%	90%
87	91%	91%	90%	88%	93%	90%	91%	90%	90%	90%	91%
88	85%	89%	89%	89%	88%	83%	86%	85%	86%	89%	87%
89	93%	95%	89%	90%	95%	90%	92%	91%	92%	92%	92%
Average for Pair	88%	89%	89%	87%	88%	87%	88%	88%	88%	88%	88%
SE	*1.4%*	*1.3%*	*1.1%*	*1.0%*	*1.4%*	*1.3%*	*1.4%*	*1.3%*	*1.3%*	*1.1%*	*1.1%*

REFERENCES

Dingus, T. A., Hankey, J., Antin, J. F., Lee, S. E., Eichelberger, L., Stulce, K., McGraw, D., Perez, M., & Stowe, L. (2015). *Naturalistic Driving Study: Technical coordination and quality control* (SHRP 2 Report S2-S06-RW-1). Washington, DC: Transportation Research Board of the National Academies.

International Organization for Standardization (ISO). (2014a). *Road Vehicles – Measurement of driver visual behaviour with respect to transport information and control systems – Part 1: Definitions and metrics* (ISO Standard 15007-1:2014(E)).

International Organization for Standardization (ISO). (2014b). *Road Vehicles – Measurement of driver visual behaviour with respect to transport information and control systems – Part 2: Equipment and procedures* (ISO Standard ISO/TS 15007-2:2014(E)).

Society of Automotive Engineers (SAE). (2000, July). *Definitions and experimental measures related to the specification of driver visual behavior using video based techniques* (SAE Standard J2396).