09-UF-004

The National Surface Transportation Safety Center for Excellence

Development and Evaluation of a Naturalistic Observer Rating of Drowsiness Protocol

Final Report

Douglas M. Wiegand • Julie McClafferty • Shelby E. McDonald • Richard J. Hanowski

Submitted: February 25, 2009

Lighting | Technology
Fatigue | Aging

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, Richard Deering from General Motors Corporation, Carl Andersen from the Federal Highway Administration (FHWA), and Gary Allen 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

Virginia Tech Transportation Institute (VTTI) researchers have developed a method for rating driver drowsiness based on the evaluation of naturalistic video footage of the driver's face and upper torso. This measure, referred to as the Observer Rating of Drowsiness (ORD; Wierwille & Ellsworth, 1994)[1] is based on subjective assessments of the driver's facial tone, behavior, and mannerisms, and is set to a 100-point continuous scale. ORD is assessed based on the 60 seconds of video prior to a trigger event (or baseline epoch). Therefore, ORD is a relatively quick/efficient method for assessing one's drowsiness level, which can then be used to describe a driver's state and investigate whether drowsiness was a contributing factor to a safety-critical event.

ORD was developed and evaluated using simulated driving data, and there has been no formal training protocol in place for data reductionists who would perform the ratings. The Wierwille and Ellsworth (1994)[1] study of ORD involved giving raters written descriptions of drowsiness levels that were to serve as a guide when making the experimental ratings. The results of this seminal study indicated that it is plausible to have a good amount of consistency within and among independent raters when assessing the level of driver drowsiness based on driver's characteristics and behaviors.

The purpose of the current study was to develop a rigorous training protocol for ORD using naturalistic driving examples. This training protocol was then evaluated in an experiment whereby the previous method for conducting ORD (reading written descriptions of drowsiness levels) was tested against the new training protocol, which includes naturalistic driving video examples and in-depth explanations of how ORD ratings are to be determined.

The training protocol was developed using video examples from both light-vehicle and heavy-vehicle naturalistic data sets as well as information regarding the definition and purpose of ORD, guidelines for performing the measures, and step-by-step instructions for completing the ratings in VTTI's Data Analysis and Reduction Tool (DART) software. Video segments were identified and reviewed by the research team to showcase a variety of relative indicators of drowsiness, including driver facial characteristics, behaviors, and mannerisms. Also, video examples were selected for individual drivers who had been identified as experiencing a wide range of drowsiness during the naturalistic study period. These examples were put in order of increasing drowsiness to present trainees with examples of how drowsiness progresses and how one can differentiate between the various levels of drowsiness. The training protocol underwent a peer review process with senior researchers at VTTI, including Dr. Walter Wierwille, who originally developed the ORD measure. The results of this peer review were positive, and all involved agreed that, anecdotally, the training protocol would improve the quality of ORD ratings in the future.

Scientific evaluation of the training protocol revealed that intra-rater reliability, inter-rater reliability and indications of validity were satisfactory. When compared against the previous methodology developed by Wierwille and Ellsworth (1994)[1], the new training protocol improved upon inter-rater reliability and indications of validity. Intra-rater reliability was stronger when using the previous methodology; however, it is speculated that this may have been due to participants more easily recognizing when a video was a repeat and thus scoring it as they

believe they had previously done. Raters who received the training tended to rate duplicate videos consistently, rated the segments consistently within their training groups, and produced scores more consistent with gold standard ratings than those who did not receive the training.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

DART	Data Analysis and Reduction Tool
DDWS FOT	Drowsy Driving Warning System Field Operational Test
ORD	Observer Rating of Drowsiness
VTTI	Virginia Tech Transportation Institute
CMV	commercial motor vehicle

CHAPTER 1. INTRODUCTION

Driver Drowsiness is a major area of concern in ground transportation safety. It is a condition which crosses all driving domains (i.e., heavy and light vehicles; commercial and private use), affects all drivers at some point, and is a contributing factor in a significant number of crashes. For example, the National Sleep Foundation's (2008)[2] "Omnibus Sleep in America Poll" found that 32 percent of those interviewed (N = 1,000) reported driving while drowsy at least once a month in the past year, while 36 percent admitted to falling asleep at the wheel in the past year.

Researchers at the Virginia Tech Transportation Institute (VTTI) conducted the "100-Car Study" which recorded naturalistic data on 100 light vehicles (241 primary and secondary drivers) over a period of 13 months, covering approximately 2 million vehicle miles of driving behavior (Dingus et al., 2006).[3] Analyses indicated that driver drowsiness was a contributing factor in 20 percent of the 82 crashes and 16 percent of the 761 near-crashes.

While driver drowsiness is prominent for all types of vehicle operators, the nature of commercial motor vehicle (CMV) operations puts these professional drivers at increased risk. CMV operators may drive up to 11 hours continuously before taking a break, often drive at night, and sometimes have irregular and unpredictable work/sleep schedules. Much of their mileage is compiled during long trips on Interstate and other divided highways. Because of their greater mileage exposure and other factors, CMV drivers' risk of being involved in a drowsiness-related crash is far greater than that of non-commercial drivers. For example, in a study of 593 randomly selected long-distance truck drivers, 47.1 percent reported having fallen asleep at the wheel of their truck at some time in their career, while 25.4 percent admitted falling asleep at the wheel in the past year (McCartt, Rohrbaugh, Hammer, & Fuller, 2000).[4]

In an investigation of 182 fatal-to-the-driver CMV crashes over a one-year period, researchers at the Transportation Safety Board (1990)[5] determined the most frequently cited probable cause was driver drowsiness (57 crashes or 31 percent). In a naturalistic study of local/short-haul truck drivers, Hanowski et al. (2000)[6] identified driver drowsiness as a contributing factor in 21 percent of 249 critical incidents. These findings were recently replicated in an analysis of 16 months of continuous naturalistic driving data with up to 103 long-haul and line-haul commercial drivers (Wiegand et al., 2008).[7] In this study, driver drowsiness was found to be a contributing factor in 16 percent of crashes and near-crashes (n = 134). Using naturalistic video data, Wiegand et al. (2008)[7] also implemented two independent measures of driver drowsiness (described in more detail below) and found that for the Observer Rating of Drowsiness (ORD) measure, drivers were rated as drowsy in 26 percent of the total safety-critical events identified (n = 952). Using another measure of drowsiness (PERCLOS, threshold of 8 percent; described more below), it was found that drowsiness was a factor in 21 percent of total safety-critical events (n = 807).

Understanding the nature of drowsiness-related safety-critical events requires a systematic approach to evaluate the entire driving situation, including driver characteristics (e.g., age), environmental parameters (e.g., road type, time of day, presence of other vehicles and other drivers' behavior), vehicle factors (e.g., vibrations); and organizational policies and practices (e.g., hours-of-service regulations; e.g., Emery & Trist, 1960).[8] Unfortunately, most drowsiness-related studies have investigated the situation after the fact, which relies heavily on

assumptions and (perhaps faulty) memory. Additionally, many past studies investigating the role of driver drowsiness in crashes are limited in the number and type of variables available for analysis (e.g., no objective measures of speed, steering wheel movement, or driver behavior before the crash). A solution to this problem is to conduct naturalistic studies, like those cited above, in which objective data on the driver, vehicle, and driving environment are recorded in real time during regular operations.

By conducting naturalistic studies, researchers can view and code safety-critical events, including observable aspects of driver errors and other behaviors which led to the events. This includes unsafe pre-event behaviors such as speeding or tailgating, as well as specific driver errors resulting in incidents.

VTTI specializes in using technology to conduct naturalistic driving studies. Technicians at VTTI equip vehicles with video cameras and other instrumentation to continuously record various performance data, driver behavior, and the driving environment. By obtaining these data, researchers can view crashes and near-crashes and associated variables/behaviors as they occur in real time, thus eliminating the need to rely on the memory of the driver or other assumptions.

VTTI researchers have developed a method for rating driver drowsiness based on the evaluation of naturalistic video footage of the driver's face and upper torso. This measure, referred to as the Observer Rating of Drowsiness (ORD; Wierwille & Ellsworth, 1994)[1] is based on subjective assessments of the driver's facial tone, behavior, and mannerisms, and is set to a 100-point continuous scale (figure 1). ORD is assessed based on the 60 seconds of video prior to a trigger event (or baseline epoch). Therefore, ORD is a relatively quick/efficient method for assessing one's drowsiness level, which can then be used to describe a driver's state and investigate whether drowsiness was a contributing factor to a safety-critical event.

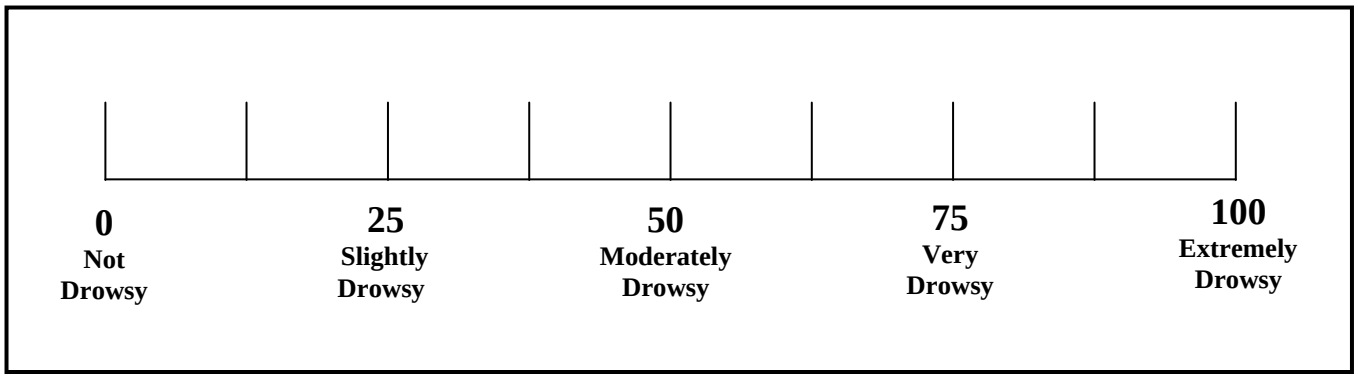


Figure 1. Diagram. ORD scale (adapted from Wierwille & Ellsworth, 1994).[1]

ORD was developed and evaluated using simulated driving data, and there has been no formal training protocol in place for data reductionists who would perform the ratings. Instead, reductionists would conduct their ratings based on written paragraph descriptions of progressive levels of drowsiness (as shown in table 1). Reductionists were instructed to use these written descriptions supplemented by their own judgment as they deemed necessary.

Table 1. Descriptions of progressive drowsiness levels.

Not Drowsy: A driver who is not drowsy while driving will exhibit behaviors such that the appearance of alertness will be present. For example, normal facial tone, normal fast eye blinks, and short ordinary glances may be observed. Occasional body movements and gestures may occur.
Slightly Drowsy: A driver who is slightly drowsy while driving may not look as sharp or alert as a driver who is not drowsy. Glances may be a little longer and eye blinks may not be as fast. Nevertheless, the driver is still sufficiently alert to be able to drive.
Moderately Drowsy: As a driver becomes moderately drowsy, various behaviors may be exhibited. These behaviors, called mannerisms, may include rubbing the face or eyes, scratching, facial contortions, and moving restlessly in the seat, among others. These actions can be thought of as countermeasures to drowsiness. They occur during the intermediate stages of drowsiness. Not all individuals exhibit mannerisms during intermediate stages. Some individuals appear more subdued, they may have slower closures, their facial tone may decrease, they may have a glassy-eyed appearance, and they may stare at a fixed position.
Very Drowsy: As a driver becomes very drowsy, eyelid closures of 2 to 3 seconds or longer usually occur. This is often accompanied by a rolling upward or sideways movement of the eyes themselves. The individual may also appear not to be focusing the eyes properly, or may exhibit a cross-eyed (lack of proper vergence) look. Facial tone will probably have decreased. Very drowsy drivers may also exhibit a lack of apparent activity, and there may be large isolated (or punctuating) movements, such as providing a large correction to steering or reorienting the head from a leaning or tilted position.
Extremely Drowsy: Drivers who are extremely drowsy are falling asleep and usually exhibit prolonged eyelid closures (4 seconds or more) and similar prolonged periods of lack of activity. There may be large punctuated movements as they transition in and out of intervals of dozing.

The purpose of the current study was to develop a rigorous training protocol for ORD using naturalistic driving examples. This training protocol was then evaluated in an experiment whereby the previous method for conducting ORD (described above with no formal training) was tested against the new training protocol. This report describes the past work on ORD, the development of the new training protocol, and the methods and results of the evaluation of the new protocol.

CHAPTER 2. PREVIOUS ORD RESEARCH

The ORD measure was first developed and evaluated by Wierwille and Ellsworth (1994)[(1)], who came up with rating system based on the observation that researchers could estimate a subject's level of drowsiness based on characteristics such as facial tone, eye closures, and behaviors (e.g., rubbing face, yawning, stretching). While the experiment is more fully described in the aforementioned journal article, it is summarized here for reference.

METHODS/DESIGN

Six graduate students in human factors engineering were chosen as participants as the experimenters believed they were familiar with rating procedures and had received behavioral training. The participants were presented with the ORD scale (figure 1) and a one-page document describing the various levels of drowsiness (see table 1). They were then presented with 27 one-minute video segments of drivers at various levels of drowsiness which were obtained from previous drowsiness studies using a moving-base driving simulator. Three of these segments were duplicates to assess intra-rater reliability. In addition, two sessions were held one week apart to assess test-retest reliability.

RESULTS

When assessing intra-rater reliability, the Pearson *r* correlation had a value of 0.88 ($t = 10.92$, $df = 34$; $p < .001$), indicating that raters consistently rated the level of drowsiness when asked to rate duplicate video segments. When assessing test-retest reliability, the Pearson *r* correlation had a value of 0.81 ($t = 5.45$, $df = 16$; $p < .001$), indicating that raters consistently rated the level of drowsiness when asked to rate the same segments twice, separated by a period of one week.

When averaging the individual Pearson r correlations of the raters to determine inter-rater reliability, the value was 0.81 ($t = 3.71$, $df = 26$; $p < .001$), indicating the ratings tended to be consistent between individual raters. The individual rater correlations ranged between 0.68 to 0.91, as shown in table 2.

Table 2. Inter-rater correlation matrix.

	Rater Number				
Rater Number	2	3	4	5	6
1	0.87	0.81	0.68	0.72	0.81
2		0.85	0.79	0.85	0.91
3			0.84	0.80	0.86
4				0.84	0.76
5					0.80

RATIONALE FOR PRESENT STUDY

The results from Wierwille and Ellsworth's (1994)[1] study indicate that it is plausible to have a good amount of consistency within and among independent raters when assessing the level of driver drowsiness based on drivers' characteristics and behaviors. Since this study, ORD has been used as a measure of driver drowsiness in several naturalistic driving studies performed by VTTI researchers. In these studies, the same methodology has been used for instructing the raters before assessing video segments (i.e., they were given the written descriptions shown in table 1). A number of raters in these experiments have commented that they lack confidence in the accuracy of their ORD ratings, often wondering if they are doing the ratings correctly.

Based on this feedback, the researchers in the present study developed an extensive training protocol to explain the rating process in detail, and to show trainees video examples of the physical characteristics, behaviors, and mannerisms believed to be relative indicators of drowsiness. In addition, video examples showing individual drivers at various levels of drowsiness were included in the protocol to show trainees the progression of drowsiness (i.e., how the same person may look at various levels of drowsiness).

The present study was performed to address the following research questions:

1. Can the Wierwille and Ellsworth (1994)[1] results be replicated using naturalistic driving data (as opposed to simulated driving data)?

2. Will implementation of a rigorous training protocol, including naturalistic driving video examples of driver characteristics and behaviors and individual driver drowsiness progressions, improve the reliability of ratings when compared to the methodology of the Wierwille and Ellsworth (1994)[1] study?

3. Is graduate coursework and experience in human factors engineering, psychology, or a related field necessary to perform ORD ratings reliably?

4. How well will naive raters (i.e., those with no prior ORD experience) perform on ORD ratings when compared to researchers who have expertise in conducting these ratings?

To answer these questions, researchers first developed a rigorous ORD training protocol, and then performed an experiment to evaluate this protocol in comparison to Wierwille and Ellsworth's (1994)[1] methodology. The sections below describe these activities.

CHAPTER 3. DEVELOPMENT OF THE ORD TRAINING PROTOCOL

NATURALISTIC DATA SOURCES FOR VIDEO EXAMPLES

Two large-scale naturalistic driving databases were used to identify video examples for the ORD training protocol. These studies include:

The Drowsy Driving Warning System Field Operational Test (DDWS FOT; Hanowski et al., 2005; Hickman et al., 2005; Wiegand, Hanowski, Olson, & Melvin, 2008).[9,10,7] This study involved the instrumentation of 46 heavy vehicles (tractor trailers) and included 103 professional truck driver participants over the course of 16 months (approximately 4-5 months of driving for each driver).

The 100-Car Naturalistic Driving Study (Dingus et al., 2006).[3] This on-road study involved 109 primary drivers (241 total drivers; primary plus secondary) of light vehicles over a period of 13 months (approximately 12 months of driving for each driver).

IDENTIFYING RELATIVE INDICATORS OF DROWSINESS

The first step in the development of the ORD training protocol was to identify the relative indicators of drowsiness (i.e., physical characteristics, behaviors, and mannerisms). VTTI researchers (n = 8) who have had experience with conducting ORD ratings in the past were asked to list as many relative indicators of drowsiness as they could think of based on their experience. Their responses were merged into a single list, which was then edited for clarification and to reduce redundancy. The final list of relative indicators of drowsiness is listed below:

- Rubbing or scratching of face, head, or neck
- Yawning
- Moving restlessly in seat (e.g., adjusting posture)
- Nodding/drooping head
- Slow eye closures, eyes rolling back
- Glassy eyes
- Squinting eyes
- Blank stare
- Fixed gaze
- Strained efforts to open eyes wide (e.g., blinking hard, then opening eyes wide)
- Leaning face in hands
- Biting/licking lips
- Stretching
- Slouching
- Leaning head back
- Loss of neck muscle control/head bobbing
- Facial tone/sagging of facial features
- Facial contortion (e.g., eyebrows)
- Shaking head rapidly
- Lack of activity

Once this list of relative indicators of drowsiness was developed, a team of three data reductionists began searching the DDWS FOT and 100-Car databases to identify video examples of each. All video examples were reviewed and evaluated by the research team, which selected what they agreed to be the best examples. If the quality of a video example was called into question, it was replaced.

A video example was identified for both the DDWS FOT and 100-Car studies for each relative indicator of drowsiness. In addition to ensuring that a heavy and light vehicle example was identified for each indicator, the researchers made every effort to include men and women, as well as drivers of various ethnicities and facial features. Video examples were then edited to show only the driver camera angle (i.e., the forward roadway, side views, etc. were removed) so the trainees would focus solely on the driver.

In addition, a behavior and mannerism checklist (figure 2) was developed for the protocol as a tool for individuals to use while performing ORD ratings. This checklist is to be used by raters to keep track of what they observe (and to what extent) during the 1 minute of video upon which ORD is based.

ORD Behavior & Mannerism Checklist

Eyes/Eyebrows:	None	Minor	Moderate	Extreme
Rubbing/ Scratching	☐	☐	☐	☐
Blank/Fixed Stare	☐	☐	☐	☐
Squinting	☐	☐	☐	☐
Excessive/Hard Blinking	☐	☐	☐	☐
Slow Closure	☐	☐	☐	☐
Unfocused rolling	☐	☐	☐	☐
Glassy/Glazed	☐	☐	☐	☐
Raise/Open Wide	☐	☐	☐	☐
Lower/Scowl	☐	☐	☐	☐
Body:				
Slumping/Slouching/ Leaning	☐	☐	☐	☐
Sighing	☐	☐	☐	☐
Stretching	☐	☐	☐	☐
Body Rolling/Slack Muscle Tone	☐	☐	☐	☐
Body Position Change (restlessness)	☐	☐	☐	☐

Mouth:	None	Minor	Moderate	Extreme
Yawning	☐	☐	☐	☐
Biting/Licking Lips	☐	☐	☐	☐
Tongue Motion	☐	☐	☐	☐
Face:				
Rubbing/Holding	☐	☐	☐	☐
Facial Contortions	☐	☐	☐	☐
Slack Muscle Tone	☐	☐	☐	☐
Neck/Head:				
Hair: Scratching/ Straightening	☐	☐	☐	☐
Rubbing/Holding	☐	☐	☐	☐
Head Leaning (back or side, unsupported	☐	☐	☐	☐
Head Position Change	☐	☐	☐	☐
Head Nodding/ Drooping	☐	☐	☐	☐

Other Notes:

Figure 2. Chart. The ORD Behavior & Mannerism Checklist.

IDENTIFYING INDIVIDUAL DRIVER DROWSINESS PROGRESSION EXAMPLES

In addition to the relative indicators of drowsiness, individual drivers' data were reviewed to select six drivers who exhibited a range of drowsiness during the DDWS FOT and 100-Car studies. Six drivers (three light-vehicle and three heavy-vehicle operators) were selected and video examples of them driving while alert, slightly drowsy, moderately drowsy, very drowsy, and extremely drowsy were identified. Each of these video clips was reviewed, evaluated, and edited in the same manner as the relative indicators of drowsiness videos, and the research team developed a written description of how each one was classified (e.g., what led to a driver being classified as "very drowsy" as opposed to "extremely drowsy"). The written descriptions were directly below the links to the video examples in the training protocol.

FINALIZING THE ORD TRAINING PROTOCOL

The finalized training protocol is available in Appendix A, and contains the following content:

- Definition and Purpose of ORD
- The ORD Continuum: The continuum was adapted somewhat from figure 1, above, to include color.
- Five Levels of Drowsiness Descriptions
- Tips for Rating Drowsiness: These were general guidelines that were developed to ensure consistent, repeatable, and unbiased ratings.
- Driver Appearance, Behaviors, and Mannerisms Indicating Drowsiness: This section includes a description plus brief video examples for light- and heavy-vehicle drivers.
- ORD Examples (Driver Progressions): This section includes an overall description of the driver progressions, the 1-minute video examples, and the written descriptions of each.
- The ORD Behavior and Mannerism Checklist (figure 2): This section includes instructions for how to use the checklist.
- Instructions for Determining and Recording ORD Ratings: This section includes step-by-step instructions for how to use VTTI's Data Analysis and Reduction Tool (DART) software to view driving events and make ORD ratings.

Once the training protocol was developed, the research team held a peer review meeting to solicit feedback from senior research faculty at VTTI regarding the protocol, video examples, and study design for evaluating the training protocol (see below). This feedback was then incorporated into the finalized protocol document.

CHAPTER 4. EVALUATING ORD TRAINING CONDITIONS

An experiment was conducted to answer the research questions stated above.

EXPERIMENTAL TASK AND DESIGN

The research team identified 24 video segments from the DDWS FOT and 100-Car studies to be rated for the experiment. These video segments represented a range of drowsiness levels, heavy- and light-vehicle drivers, men and women, and drivers of various ethnicities and facial characteristics. Three of the video segments were repeated to assess intra-rater reliability, thus participants were asked to rate 27 videos total. Participants' ratings of the videos constituted the dependent variable in this study.

The independent variables in this experiment were the training condition assigned and the educational level of the raters. The three conditions were:

Group 1 (n = 8): Individuals with at least some graduate coursework in human factors engineering, psychology, or a related field who were asked to review the written descriptions of the five levels of drowsiness (table 1) and were given instructions for completing the 27 ORD ratings in DART.

Group 2 (n = 8): Individuals with at least some graduate coursework in human factors engineering, psychology, or a related field who attended a ~2-hour training session in which the ORD training protocol developed for this study was reviewed/explained before the participants completed the 27 ORD ratings in DART.

Group 3 (n = 8): Undergraduate data reductionists who attended a ~2-hour training session in which the ORD training protocol developed for this study was reviewed/explained before the participants completed the 27 ORD ratings in DART. The training session was identical to that in Group 2; the difference between the groups being the educational level of the participants.

The three experimental groups met independently in a conference room at VTTI to receive instructions and training, if applicable. Participants were discouraged from asking questions or creating discussion to control for potential bias. Following the instructions/training, participants relocated to a data reduction lab, where they rated the experimental events in DART. Total experimental time was approximately 2 hours for Group 1, and 4.5 hours for Groups 2 and 3.

To control for presentation effects, a balanced Latin square design was used to determine the order in which participants viewed and rated the experimental events. Each participant was given a sheet of paper with the list of experimental events in the order in which they were to rate them. Participants were told not to deviate from the order presented to them.

RESULTS

Several sets of analyses, including Pearson *r* correlations and paired sample t-tests, were performed to test intra-rater reliability (how consistently individual participants rated repeated events), inter-rater reliability (how consistent ratings were across participants), and how these

results compared across experimental groups/conditions. In addition, analyses were performed to determine an indication of the validity of ratings based on comparisons with "gold standard" ratings performed by three members of the research team who are considered experts at ORD. The results of these analyses are discussed below.

Intra-rater Reliability

Three of the video segments were duplicated per rater for each of the experimental conditions to assess intra-rater reliability.

Table 3 shows the Pearson *r* correlations for each of the pairings for the Group 1 raters (graduate/post-graduate researchers who did not receive the new training).

Table 3. Group 1 correlations for video segment duplicates.

Video Segment Duplicates	Pearson *r*	Significance
Pair 1 (n = 8)	0.94	0.001
Pair 2 (n = 8)	0.92	0.001
Pair 3 (n = 8)	0.83	0.012

In addition, the absolute values of differences in first exposure to second exposure of each duplicate pairing were calculated. Table 4 shows the results of a t-test on these data across all three duplicate pairs for Group 1. While the t-test shows that the mean difference was significantly different from zero, a mean difference of 5.21 is still practical since 5 points on a 100-point scale would not necessarily constitute a different interpretation of drowsiness level.

Table 4. Group 1 T-test on absolute differences on duplicate exposures of video segments.

Group 1	*t*	df	Sig	Mean Diff	Std Dev	Lower CI	Upper CI
	3.22	23	0.004	5.21	7.94	1.86	8.57

Table 5 shows the Pearson *r* correlations for each of the pairings for the Group 2 raters (graduate/post-graduate level researchers who received the new training).

Table 5. Group 2 correlations for video segment duplicates.

Video Segment Duplicates	Pearson *r*	Significance
Pair 1	0.64	0.087
Pair 2	0.51	0.192
Pair 3	0.97	0.000

As with Group 1, the absolute values of differences in first exposure to second exposure of each duplicate pairing were calculated for Group 2. Table 6 shows the results of a t-test on these data across all three duplicate pairs for Group 2. While the t-test shows that the mean difference was significantly different from zero, a mean difference of 11.27 is still practical since 11 points on a 100-point scale would not necessarily constitute a different interpretation of drowsiness level.

Table 6. Group 2 T-test on absolute differences on duplicate exposures of video segments.

Group 2	*t*	df	Sig	Mean Diff	Std Dev	Lower CI	Upper CI
	4.142	23	0	11.27	13.33	5.64	16.9

Table 7 shows the Pearson *r* correlations for each of the pairings for the Group 3 raters (undergraduate data reductionists who received the new training).

Table 7. Group 3 correlations for video segment duplicates.

Video Segment Duplicates	Pearson *r*	Significance
Pair 1	0.52	0.184
Pair 2	0.75	0.034
Pair 3	0.03	0.944

One's attention may be drawn to the extremely low correlation for Pair 3 in this table. Upon examining the data more closely, two outliers were detected in which one rater had a range of 58 points for the two videos, while another had a range of 40 points. It is unclear why this is the case. To rate the same video segment in opposite extremes after repeated exposure is unusual, and it may be the case that it was unintentional. When these outliers are removed, the correlation rises to $r = 0.82$ ($p = .045$).

Again, the absolute values of differences in first exposure to second exposure of each duplicate pairing were calculated for Group 3. Table 8 shows the results of a t-test on these data across all three duplicate pairs for Group 3. While the t-test shows that the mean difference was significantly different from zero, a mean difference of 11.53 is still practical since 11-12 points on a 100-point scale would not necessarily constitute a different interpretation of drowsiness level.

Table 8. Group 3 T-test on absolute differences on duplicate exposures of video segments.

Group 3	*t*	df	Sig	Mean Diff	Std Dev	Lower CI	Upper CI
	3.84	23	0.001	11.53	14.7	5.32	17.73

When removing the two outliers as described above, the results for Group 3 change favorably to that seen in table 9. The mean difference when the outliers are removed falls to 8.13, which again is practical given the nature of the measure.

Table 9. Group 3 T-test on absolute differences on duplicate exposures of video segments (outliers removed).

Group 3	*t*	df	Sig	Mean Diff	Std Dev	Lower CI	Upper CI
	4.17	21	0.000	8.13	9.16	4.07	12.19

Inter-rater Reliability

To evaluate inter-rater reliability, the average of each group's rating per video segment was calculated, and then individual ratings were correlated with this average. The overall correlation value for inter-rater reliability (as determined by the mean of the individual Pearson r correlations) was 0.72 for Group 1, indicating the ratings tended to be consistent between individual raters. The individual rater correlations ranged between 0.50 to 0.88, as shown in table 10.

Table 10. Group 1 inter-rater correlation matrix.

	Rater Number						
Rater Number	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0.76	0.85	0.50	0.77	0.69	0.76	0.53
2		0.79	0.84	0.72	0.80	0.72	0.66
3			0.63	0.85	0.76	0.86	0.65
4				0.68	0.72	0.62	0.55
5	Avg *r* = 0.72				0.80	0.88	0.63
6						0.68	0.74
7							0.64

The overall correlation value for inter-rater reliability (as determined by the mean of the individual Pearson r correlations) was 0.74 for Group 2, indicating the ratings tended to be consistent between individual raters. The individual rater correlations ranged between 0.54 to 0.84, as shown in table 11.

Table 11. Group 2 inter-rater correlation matrix.

	Rater Number						
Rater Number	**10**	**11**	**12**	**13**	**14**	**15**	**16**
9	0.64	0.71	0.66	0.72	0.64	0.66	0.68
10		0.78	0.78	0.78	0.73	0.80	0.74
11			0.84	0.81	0.68	0.81	0.77
12				0.84	0.54	0.71	0.77
13	Avg r = 0.74				0.76	0.80	0.81
14						0.73	0.72
15							0.73

The overall correlation value for inter-rater reliability as determined by the mean of the individual Pearson r correlations was 0.77 for Group 3, indicating the ratings tended to be consistent between individual raters. The individual rater correlations ranged between 0.69 to 0.90, as shown in table 12.

Table 12. Group 3 inter-rater correlation matrix.

	Rater Number						
Rater Number	**18**	**19**	**20**	**21**	**22**	**23**	**24**
17	0.76	0.79	0.88	0.85	0.77	0.70	0.79
18		0.72	0.84	0.81	0.76	0.71	0.69
19			0.78	0.73	0.72	0.72	0.81
20				0.86	0.82	0.71	0.79
21	Avg r = 0.77				0.84	0.72	0.82
22						0.58	0.90
23							0.70

Paired sample t-tests were also performed to determine whether the average Pearson r was significantly different between groups. Table 13 shows the results of these tests, which indicate the only statistically significant difference between groups was between Groups 1 and 3 ($p < .05$).

Table 13. Paired T-tests of group inter-rater correlation means.

	Mean	Std. Dev.	t	df	Sig
Group 1	0.72	0.10	-0.94	27	0.358
Group 2	0.74	0.07			
Group 1	0.72	0.10	-2.15	27	0.041
Group 3	0.77	0.07			
Group 2	0.74	0.07	-1.74	27	0.093
Group 3	0.77	0.07			

Indication of Validity

To determine an indication of validity for the measure, three members of the research team who are considered experts in ORD rated the video segments. For each video segment, the three expert raters' average was calculated to use as a "gold standard" rating for each segment. These gold standard ratings were then used as indicators of validity to which the study participants' ratings could be compared to. Each expert rater's scores for the video segments, as well as the gold standard rating, are available in Appendix B.

To determine how different the participants' ratings were from the gold standard ratings (i.e., error), the participants' ratings were subtracted from the gold standard rating, and converted into absolute values. The group means of absolute error were then calculated, and paired sample t-tests were performed to make comparisons between groups. Table 14 shows the results of these tests, which shows that each comparison of means was statistically different. Overall, Group 3 had the least amount of error based on the gold standard ratings, followed by Group 2 and then Group 1.

Table 14. Paired T-tests of mean absolute error in comparison to gold standard ratings.

	Mean	Std. Dev.	t	df	Sig
Group 1	24.84	18.99	3.3	191	0.001
Group 2	20.20	16.70			
Group 1	24.84	18.99	5.57	191	0
Group 3	16.92	14.93			
Group 2	20.20	16.70	2.53	191	0.012
Group 3	16.92	14.93			

CHAPTER 5. DISCUSSION

This project involved the development and evaluation of a naturalistic ORD training protocol. The protocol was developed using video examples from both light-vehicle and heavy-vehicle naturalistic data sets as well as information regarding the definition and purpose of ORD, guidelines for performing the measures, and step-by-step instructions for completing the ratings in VTTI's DART software. Video segments were identified and reviewed by the research team to showcase a variety of relative indicators of drowsiness, including driver facial characteristics, behaviors, and mannerisms. Also, video examples were selected for individual drivers who had been identified as experiencing a wide range of drowsiness during the naturalistic study period. These examples were put in order of increasing drowsiness to present trainees with examples of how drowsiness progresses, and how one can differentiate between the various levels of drowsiness shown in table 1. The training protocol underwent a peer review process with senior researchers at VTTI, including Dr. Walter Wierwille, who originally developed the ORD measure. The results of this peer review were positive, and all involved agreed that, anecdotally, the training protocol would improve the quality of ORD ratings in the future.

To evaluate the training protocol scientifically, an experiment was conducted which tested three separate groups of raters: graduate/post-graduate level researchers who only received the written descriptions of drowsiness (table 1) and did not receive the new training; graduate/post-graduate level researchers who received the new training; and undergraduate level data reductionists who received the new training.

Three research questions were posed at the beginning of the study. These questions are addressed below.

Can the Wierwille and Ellsworth (1994)[1] results be replicated using naturalistic driving data (as opposed to simulated driving data)?

As described above, Wierwille and Ellsworth (1994)[1] evaluated the original ORD scale and drowsiness level descriptions using graduate level researchers who performed the ratings on simulated driving data. This study resulted in an intra-rater reliability of $r = 0.88$ and an average inter-rater reliability of $r = 0.81$, which indicated that it is plausible to have a good amount of consistency within and among independent raters when assessing the level of driver drowsiness based on drivers' characteristics and behaviors.

One experimental condition in the present study mimicked Wierwille and Ellsworth's methodology with the exception of using naturalistic driving data instead of simulated driving data (Group 1; graduate/post-graduate level researchers who reviewed the written descriptions of drowsiness levels but did not receive the newly developed training). In terms of intra-rater reliability, this group had excellent correlations between first and second exposure to duplicate videos, with Pearson r values ranging from 0.83 – 0.94. This is comparable to what Wierwille and Ellsworth had found. In terms of inter-rater reliability, the average inter-rater correlation for this group was $r = 0.72$, which is somewhat weaker than that reported by Wierwille and Ellsworth. Nonetheless, a correlation of above 0.70 is considered acceptable in terms of reliability (Nunnally, 1978; Pedhazur & Schmelkin, 1991).[11,12] In addition, though the t-test on the absolute differences between ratings indicated that the differences were significantly

different from zero, the mean difference was 5.21 on a 100-point scale, which is not a practically significant difference. With this said, the results of Group 1 indicate that the results of Wierwille and Ellsworth's (1994)[1] seminal work can be replicated using naturalistic driving data.

Will implementation of a rigorous training protocol, including naturalistic driving video examples of driver characteristics and behaviors and individual driver drowsiness progressions, improve the reliability of ratings when compared to the methodology of the Wierwille and Ellsworth (1994)[1] study?

To answer this question, one can compare the results of Group 1, as described above, to those of Group 2, which included graduate/post-graduate level researchers who underwent the training developed for this project. In terms of intra-rater reliability, Group 2's Pearson r values ranged from 0.51 - .097, which is a larger range than that of Group 1, and includes values which are less than the acceptability threshold of 0.70. Nonetheless, the absolute differences between ratings averaged to 11.27, which is not necessarily a practically significant difference with the 100-point ORD scale.

The amount of variance in intra-rater reliability between Group 1 and Group 2 is surprising. It is possible that since Group 2 raters had received the training (and were therefore more attuned to specific behavioral indicators of drowsiness) they noticed different driver characteristics and mannerisms between ratings of the same video. It also may be the case that since Group 1 did not have the task of completing the ORD Behavior and Mannerism Checklist (shown in figure 2), that raters completed the ratings more quickly and therefore may have more easily noticed the similarity in repeated videos, thus assigning what they believed to be a consistent rating with their first rating of that video segment.

In terms of inter-rater reliability, Group 2 performed slightly better than Group 1, with an average inter-rater correlation of $r = 0.74$. However, this difference was not significantly different from Group 1's correlation of $r = 0.72$ ($t = -0.94$; $p = 0.358$).

These results indicate that the training group (Group 2) performed relatively poorly on intra-rater reliability when compared to the no-training group (Group 1), and both groups were comparable in terms of inter-rater reliability, as each performed at an acceptable level. This finding is somewhat counterintuitive in that it was expected that the training would result in improved reliability. One possible explanation for the lack of improvement in reliability following training is the length of the experimental sessions. Group 1 did not receive the training, which took approximately 2 hours. Therefore, their experimental session was much shorter than Group 2's. It is possible that Group 2 participants were relatively less focused by the time they began rating the video segments, given the fact that they had been away from their regular work for a longer period of time and had just participated in a 2-hour training session. However, this is merely speculative. A post-experimental survey would have been helpful in determining if this was the case.

Is graduate coursework and experience in human factors engineering, psychology, or a related field necessary to perform ORD ratings reliably?

This question was addressed by comparing the data from Group 2 (graduate/post-graduate level researchers who underwent the training developed for this project) and Group 3 (undergraduate data reductionists who underwent the same training). In terms of intra-rater reliability, Group 3 had Pearson correlations ranging from 0.03 – 0.75 on the three repeated video segments. This indicates that, overall, Group 3 raters were much less consistent in rating the duplicate video segments when compared to both Groups 1 & 2. It is interesting that Group 1 could perform so well in terms of being consistent when rating duplicate events (again, *r*'s ranging from 0.83 - 0.94), while the two groups who received training performed much less consistently. As mentioned above, it may be the case that Group 1 was able to navigate the rating process more quickly, and may have noticed when videos were duplicated, thus assigning a similar score without treating the video segment as independent from those previously viewed. Again, a post-experimental survey would have been helpful in determining whether this was the case.

At first glance, the intra-rater reliability for Group 3 seems rather inconsistent; however, after removing several unusual outliers, this group was nearly identical to Group 2. The average correlations (when the outliers are removed) are $r = 0.70$ for Group 3 and $r = 0.71$ for Group 2. Again, these are considered acceptable indications of reliability.

Group 3 was more consistent than the others in terms of inter-rater reliability (average inter-rater correlation of $r = 0.77$). This correlation is comparable to that of Group 2 ($r = 0.74$), however it is significantly greater than that of Group 1 ($r = 0.72$). Two things may account for this difference. Group 3 consisted of students and recent college graduates whose job duties as data reductionists primarily entail analyzing naturalistic driving data, as opposed to Groups 1 & 2, whose job duties are diverse and may or may not involve any exposure to reducing naturalistic driving data. Given this, combined with the fact that Group 3 underwent the training protocol developed as part of this project, may explain why they were more consistent as a group when compared to Group 1, who received no such training.

How well will naive raters (i.e., those with no prior ORD experience) perform on ORD ratings when compared to researchers who have expertise in conducting these ratings?

Three members of the research team who are considered experts in ORD rated the experimental video segments to develop gold standard ratings to use as indicators of validity. The term *indicator of validity* is used because the expert ratings were still subjective, and a true, objective measure of drowsiness (e.g., electroencephalography) was not used. However, the gold standard ratings developed for this study are based on expert input and therefore may be considered an acceptable comparison.

Paired T-tests were performed to compare the mean absolute error from gold standard ratings between groups. These analyses revealed that Group 1 had the greatest mean absolute error ($M = 24.84$), followed by Group 2 ($M = 20.20$) and Group 3 ($M = 16.92$). When comparing these averages, Group 2 performed significantly better than Group 1 ($t = 3.3$; $p < 0.001$), indicating that the training may have impacted the validity of the ORD ratings (even though it did not improve intra-rater reliability). What is interesting, however, is that Group 3 performed

significantly better than both Group 1 ($t = 5.57$; $p < 0.001$) and Group 2 ($t = 2.53$; $p < 0.05$). When compared to Group 1, this may again be an indication that the training improved the validity of the ORD ratings. However, it is interesting that Group 3 outperformed Group 2, who had received the same training. As mentioned above, this may be due to the fact that Group 3 consisted of individuals who are much more exposed on a regular basis to naturalistic driving data than those in Groups 1 and 2. It is important to note that while those in Group 3 had been exposed to naturalistic driving data before, they had no previous experience with conducting ORD ratings (nor did any study participants). One limitation to this study is that a comparison group of undergraduate data reductionists who *did not* receive the training was not tested. Comparing this group with Group 3 would have been useful for determining whether exposure to naturalistic driving data impacted the ORD ratings of these individuals as opposed to the training session.

CHAPTER 6. CONCLUSIONS

A rigorous ORD training protocol was developed for this project. Evaluation of the training protocol revealed that intra-rater reliability, inter-rater reliability, and indications of validity were satisfactory. When compared against the previous methodology developed by Wierwille and Ellsworth (1994)[1], the new training protocol improved upon inter-rater reliability and indications of validity. Intra-rater reliability was stronger when using the previous methodology; however, it is speculated that this may have been due to participants more easily recognizing when a video was a repeat and thus scoring it as they believe they had previously done. Raters who received the training tended to rate duplicate videos consistently, rated the segments consistently within their training groups, and produced scores more consistent with gold standard ratings than those who did not receive the training.

It is recommended that the protocol developed in this project continues to be used as a training tool for data reductionists who will perform ORD ratings. While the protocol underwent a peer review process, participants in the study who were charged with the task of completing the training and ratings were not given the opportunity to provide their input/feedback regarding the process. It is assumed that the protocol could be improved given user feedback.

In addition, intra-rater and inter-rater reliability, as well as indicators of validity, could be improved upon if trainees are allowed to hold discussions and ask questions during the training session. Finally, the gold standard ratings developed for this study could be used to implement a proficiency test for trainees. If an unacceptable amount of error is found between a trainee's rating and a gold standard rating, the trainer could discuss the video segment with the trainee and retest until he or she performs better. However, it is recommended that training video segments and their respective gold standard ratings are updated frequently to eliminate potential practice effects. A protocol for establishing rater proficiency to address some of the items mentioned above is included in Appendix C.

APPENDIX A: NATURALISTIC ORD TRAINING PROTOCOL

(NOTE: PROTOCOL INCLUDES LINKS TO VIDEO EXAMPLES WHICH ARE ONLY ENABLED WHEN CONNECTED TO THE VTTI COMPUTER NETWORK)

Observer Rating of Drowsiness (ORD)

Rater Training Manual

Definition and Purpose of ORD

The Observer Rating of Drowsiness (ORD) is a subjective assessment of how drowsy a naturalistic driving study participant is based on his/her physical appearance, behaviors and mannerisms. ORD is assessed based on the 60 seconds of video prior to a trigger event (or baseline epoch). Therefore, ORD is a relatively quick/efficient method for assessing one's drowsiness level, which can then be used to describe a driver's state and investigate whether drowsiness was a contributing factor to a safety-critical event.

The ORD Continuum

The ORD Continuum is a continuous scale ranging from *Not at all Drowsy* (0) to *Extremely Drowsy* (100). Between these extremes are three levels of drowsiness: *Slightly Drowsy*, *Moderately Drowsy*, and *Very Drowsy*. Each of these five levels of drowsiness is described in more detail below.

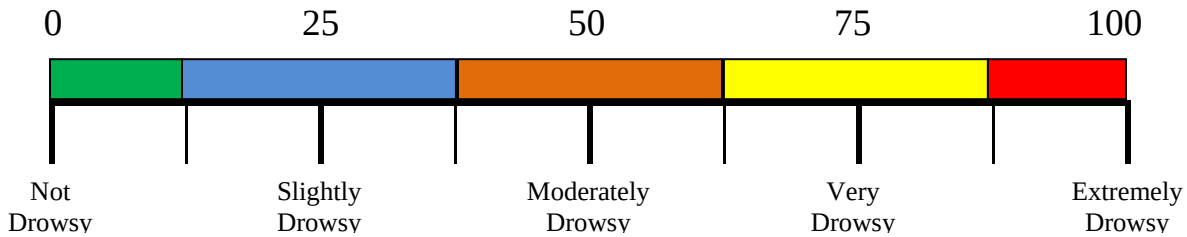
Each ORD rating is based on a review of 60 seconds of video prior to a trigger and represents **the overall or average level of drowsiness** observed over that time period. Ratings should be based on the descriptions below. However, keep in mind that these descriptions are provided as guides only, and may not completely describe what is observed in the video. If the rater believes that the descriptions overlook something important or do not properly describe what is observed, then these descriptions should be supplemented with the best judgment of the rater.

Once the appropriate video(s) have been reviewed, the ORD analyst determines where the individual being observed is on the continuum based on observable indicators of drowsiness and alertness. Raters may determine their ratings <u>at any point on the continuum</u> (i.e., any number between 0-100), not just at one of the descriptors or vertical marks.

A picture of this continuum is provided below. This continuum and the descriptions that follow were adapted from those in Wierwille and Ellsworth (1994)[1].

---

[1] Wierwille, W. W., and Ellsworth, L.A. (1994). "Evaluation of driver drowsiness by trained raters." <u>Accident Analysis and Prevention</u> **26**(5): 571-581.



Five Levels of Drowsiness

- **Not Drowsy (0 - 12.49):** A driver who is not drowsy while driving will exhibit behaviors such that the appearance of alertness will be present. For example, normal facial tone, normal fast eye blinks, and short ordinary glances may be observed. Occasional body movements and gestures may occur.

- **Slightly Drowsy (12.5 – 37.49):** A driver who is slightly drowsy while driving may not look as sharp or alert as a driver who is not drowsy. Glances may be a little longer and eye blinks may not be as fast. Nevertheless, the driver is still sufficiently alert to be able to drive.

- **Moderately Drowsy (37.5 – 62.49):** As a driver becomes moderately drowsy, various behaviors may be exhibited. These behaviors, called mannerisms, may include rubbing the face or eyes, scratching, facial contortions, and moving restlessly in the seat, among others. These actions can be thought of as countermeasures to drowsiness. They occur during the intermediate stages of drowsiness. Not all individuals exhibit mannerisms during intermediate stages. Some individuals appear more subdued, they may have slower closures, their facial tone may decrease, they may have a glassy-eyed appearance, and they may stare at a fixed position.

- **Very Drowsy (62.5 – 87.49):** As a driver becomes very drowsy, eyelid closures of 2 to 3 seconds or longer usually occur. This is often accompanied by a rolling upward or sideways movement of the eyes themselves. The individual may also appear not to be focusing the eyes properly, or may exhibit a cross-eyed (lack of proper vergence) look. Facial tone will probably have decreased. Very drowsy drivers may also exhibit a lack of apparent activity, and there may be large isolated (or punctuating) movements, such as providing a large correction to steering or reorienting the head from a leaning or tilted position.

- **Extremely Drowsy (87.5 – 100):** Drivers who are extremely drowsy are falling asleep and usually exhibit prolonged eyelid closures (4 seconds or more) and similar prolonged periods of lack of activity. There may be large punctuated movements as they transition in and out of intervals of dozing.

Tips for Rating Drowsiness

In addition to the descriptions provided above, there are several guidelines that, when followed, help ensure consistent, repeatable, and unbiased ratings.

1. When performing ORD, do not take the time of day or ambient lighting conditions (e.g.., day vs. night) into consideration. ORD ratings should be independent of these conditions because drowsiness can occur at any time of day, and in any lighting condition.
2. ORD is performed for the 60 second period preceding an identified event or point in time (i.e., trigger). When a full 60 seconds of video is available during this time period, then evaluate the entire 60 seconds. If less than 60 seconds is available, then use whatever is available, as long as it is at least 30 seconds. DO NOT perform ORD when there is less than 30 seconds of video available prior to the trigger.
3. ORD ratings should be the average condition over the time period being evaluated (e.g., 60 seconds). If the apparent level of drowsiness appears to change during the 60 seconds reviewed, provide an ORD score that you feel reflects an overall average. For example, a drowsy driver staring blankly ahead may become more alert when preparing to exit the Interstate due to mirror checks, etc.
4. Distractions and driver secondary tasks should not be used as indicators of alertness, per se. However, the secondary tasks that a driver engages in may impact other indicators of drowsiness (e.g., rate of slow eye closures or visual scanning). For example, a driver who is adjusting the radio or talking with a passenger with quick glances and body movements may have a lower ORD rating compared to a driver who engages in these activities but has slow, large, and/or punctuated movements. Remember to examine the presence/absence of drowsiness-related mannerisms, not the presence/absence of secondary tasks.
5. It is helpful (and usually necessary) when beginning ORD on a new driver to become familiar with the driver by watching a variety of unrated video examples to get a feel for how the driver appears and behaves. The helps the rater become familiar with that driver's normal vs. drowsy appearances, behaviors and mannerisms so that events can be rated appropriately. It is important to understand that different drivers within the same level of drowsiness may not display the same indicators.

Driver Appearance, Behaviors & Mannerisms Indicating Drowsiness

The list below provides a sample of appearances, behaviors, and mannerisms which are reportedly linked to drowsiness. This list is not necessarily comprehensive, yet provides a foundation for basic relative indicators of drowsiness. The links provided next to each indicator will open a video demonstrating that particular appearance/behavior/mannerism (often with other indicators demonstrated simultaneously; however, the indicator identified by the video title is the most obvious in the linked video clip). Please review this list and the corresponding videos, taking notes if desired. When reviewing the video to determine ORD, consider the number of

indicators present, the frequency with which indicators are observed (e.g., number of yawns), and the severity of indicator occurrences.

Appearance/Mannerism:	Video Examples	Trainee Notes:
Rubbing or scratching of face, head, or neck	Truck	
	Light Vehicle	
Yawning	Truck	
	Light Vehicle	
Moving restlessly in seat (e.g. adjusting posture)	Truck	
	Light Vehicle	
Nodding/drooping head	Truck	
	Light Vehicle	
Slow eye closures, eyes rolling back	Truck	
	Light Vehicle	
Glassy eyes	Truck	
	Light Vehicle	
Squinting eyes	Truck	
	Light Vehicle	
Blank stare/staring at a fixed position	Truck	
	Light Vehicle	
Fixed gaze	Truck	
	Light Vehicle	
Strained efforts to open eyes wide (e.g., blinking hard)	Truck	

then opening eyes wide)	Light Vehicle	
Leaning face in hands	Truck	
	Light Vehicle	
Biting/licking lips	Truck	
	Light Vehicle	
Stretching	Truck	
Slouching	Truck	
	Light Vehicle	
Leaning head back	Truck	
	Light Vehicle	
Loss of neck muscle control/Head bobbing	Truck	
	Light Vehicle	
Facial tone/sagging of facial features	Truck	
	Light Vehicle	
Facial contortion (e.g., eyebrows)	Light Vehicle	
Shaking head really fast	Truck	
Lack of Activity	Truck	
	Light Vehicle	

ORD Examples (Driver Progressions)

As noted earlier, different drivers may exhibit different combinations and degrees of drowsiness indicators. ORD raters should become familiar with each driver by viewing a variety of unrated events/files for that particular individual before any ORD ratings are made. This allows the rater to become familiar with that driver's normal vs. drowsy appearance, behavior, and mannerisms.

Below are six drowsiness "progressions" for individual drivers (3 truck drivers, 3 light vehicle drivers). These video examples begin when the driver is relatively alert and progress through the various levels of drowsiness. As you view these files, you will see how the appearances, behaviors and mannerisms vary between drivers. Some drivers exhibit some of the normal drowsiness indicators even when alert, and some drivers will have behaviors in response to high levels of drowsiness that at other times would be indicators of alertness. View the following video clips in order, along with the descriptions of why each clip was rated at the specified certain level.

Example Type	**Video Examples**	**Trainee Notes:**
Truck Example 1 (Note: This example does not include an "Extremely Drowsy" clip because the driver was never observed at this level)	Not at All Drowsy	
	Slightly Drowsy	
	Moderately Drowsy	
	Very Drowsy	

Truck Example 1 Descriptions

- *Not at All Drowsy:* The driver frequently scans the environment, and his eyes are almost constantly moving to check mirrors, instrument panel, etc. Some touching of the face, but overall looks alert.
- *Slightly Drowsy:* The driver has some sagging in his facial tone and he does not appear to be scanning his environment as much as a totally alert driver would be doing. Chewing food/gum sometimes is a relative indicator of drowsiness. Although these indicators are present, they are not relatively severe enough to justify a moderately drowsy rating.
- *Moderately Drowsy:* In this clip the driver is frequently touching his face, which is a relative indicator of moderate drowsiness. He also sighs and shifts around in his seat, at one point leaning his head back on the head rest. He also has evidence of slow eye closures, and his eyes look heavy at times.
- *Very Drowsy:* The driver is touching his face and rubbing his eyes quite a bit in the beginning of the clip. His eyes are heavy and his eye closures are somewhat slow. He looks as though he is trying to wake himself up.
- *Extremely Drowsy:* (none)

Example Type	**Video Examples**	**Trainee Notes:**
Truck Example 2	Not at All Drowsy	
	Slightly Drowsy	
	Moderately Drowsy	
	Very Drowsy	
	Extremely Drowsy	

Truck Example 2 Descriptions

- *Not at All Drowsy:* The driver seems to be in a good mood, laughing and talking in the beginning of the clip which indicates alertness (facial features are not sagging, but the opposite). Some touching of the face, but overall looks alert.
- *Slightly Drowsy:* The driver's gaze settles on the forward view, and he is not scanning his

environment as much as a completely alert person would. His mouth hangs open slightly which is an indicator that he is not as sharp as a completely alert driver. However, the driver is not having slow eye closures or displaying many mannerisms (e.g., touching face) which would classify him as moderately drowsy.

- *Moderately Drowsy:* The driver is not scanning his environment as frequently as an alert person would. His gaze is fixed forward with occasional glances to his instrument panel. He is smoking a cigarette, which may be an attempt to stimulate/wake himself. He also appears to bounce in his seat, which is an indication that he has loose muscles.
- *Very Drowsy:* The driver is blinking frequently, and his mouth is hanging open. He has several slow eye closures, and when he checks his mirrors his movement is slow and it seems as though his neck muscle control has decreased. He is not quite dozing off or having prolonged slow eye closures to warrant being classified as extremely drowsy.
- *Extremely Drowsy:* The driver is struggling to stay awake in the beginning of this clip. His head is bobbing (poor neck muscle control) and he has slow eye closures, sometimes lasting several seconds. His mouth is hanging open and his gaze looks dim. He appears to manipulate the radio, as if he is looking for something to keep him entertained/awake.

Example Type	**Video Examples**	**Trainee Notes:**
Truck Example 3	Not at All Drowsy	
	Slightly Drowsy	
	Moderately Drowsy	
	Very Drowsy	
	Extremely Drowsy	

Truck Example 3 Descriptions

- *Not at All Drowsy:* The driver's face is not showing any signs of sagging facial tone, and his blinks are rapid. He is carrying on a conversation on the radio and is frequently checking his driving environment.
- *Slightly Drowsy:* In this clip the driver's face is sagging somewhat (he has bags under his eyes), and he is not scanning his environment as much as a completely alert person would. He is smoking a cigarette, which may be an attempt to stimulate/wake himself.
- *Moderately Drowsy:* The driver rubs his face and head and widens his eyes as he is doing this, as if he is trying to wake himself up. His gaze is relatively fixed and he has at least one slow eye closure. These are not behaviors of someone who is alert or even slightly drowsy; however, they would need to be more severe to warrant "very drowsy".
- *Very Drowsy:* The driver's gaze is fixed and his mouth is hanging open. His facial tone is sagging more than usual, and he has periods where his brow is raised as if he is struggling to keep his eyes open. He has several slow eye closures and rubs his head/face to wake himself up. He is smoking a cigarette, which is common to this driver, but he is leaning on the window (as opposed to sitting up straight) and taking slow drags, which may be an indicator of drowsiness for this driver.
- *Extremely Drowsy:* The driver's face is sagging and he shakes his head quickly in the beginning of the clip, which is a sign that he is attempting to wake himself up. His gaze is relatively fixed, and he has slow eye closures and rubs his head and face frequently.

He stretches at one point, too, which is an indicator of drowsiness. This driver seems to be on the verge of falling asleep.

Example Type	Video Examples	Trainee Notes:
Light Vehicle Example 1	Not at All Drowsy	
	Slightly Drowsy	
	Moderately Drowsy	
	Very Drowsy	
	Extremely Drowsy	

Light Vehicle Example 1 Descriptions

- *Not at All Drowsy:* The driver has no slack in facial tone, and blinks rapidly. She seems to touch her face at times, but it is eating-related behavior, and does not necessarily indicate drowsiness.
- *Slightly Drowsy:* The driver's face is sagging and her eyes appear somewhat heavy, showing signs of drowsiness. However, she is talking on the phone and seems to be holding a steady paced conversation, which indicates some degree of alertness.
- *Moderately Drowsy:* The driver's face is sagging and she has several slow eye closures, thus being considered at least moderately drowsy. She also yawns, which is an indicator of drowsiness. If these indicators were more frequent, she would be scored higher.
- *Very Drowsy:* The driver has several slow eye closures and moments where it appears her neck muscles are not adequately supporting her head. Between slow eye closures, she is checking her environment, and she seems to be talking to herself or a passenger. Not enough slow eye closures to be considered extremely drowsy.
- *Extremely Drowsy:* The driver is struggling to stay awake. She has numerous slow eye closures lasting several seconds each. Her eyes look very heavy. Her movements seem slow (e.g., chewing gum).

Example Type	Video Examples	Trainee Notes:
Light Vehicle Example 2	Not at All Drowsy	
	Slightly Drowsy	
	Moderately Drowsy	
	Very Drowsy	
	Extremely Drowsy	

Light Vehicle Example 2 Descriptions

- *Not at All Drowsy:* The driver looks alert, checking his surroundings and talking/singing. He is blinking at a normal rate.
- *Slightly Drowsy:* The driver looks somewhat alert, though he has slight facial sagging (light circles under his eyes).
- *Moderately Drowsy:* The driver is shifting in his seat and rubbing his arms/torso, which are indicators of drowsiness. He widens his eyes quickly on several occasions, which may be a sign that he is trying to stimulate/wake himself. His gaze is somewhat fixed toward the end, so he is not monitoring the environment as much as an alert person would.

- *Very Drowsy:* The driver's face is sagging, and he has dark circles under his eyes. He is beginning to have slow eye closures. He touches his face several times and also bites/chews his lip. He takes drinks of a caffeinated beverage.
- *Extremely Drowsy:* The driver's face is sagging, and he has dark circles under his eyes. He has several slow eye closures and yawns. He looks to be struggling to keep his eyes open. He shakes his head and touches his head/hair/face, and seems to have a slight loss of control of his neck at times.

Example Type	Video Examples	Trainee Notes:
Light Vehicle Example 3	Not at All Drowsy	
	Slightly Drowsy	
	Moderately Drowsy	
	Very Drowsy	
	Extremely Drowsy	

Light Vehicle Example 3 Descriptions

- *Not at All Drowsy:* The driver scans his environment and is sitting straight up. He is chewing gum, but he seems to be doing it quickly. Overall he looks alert.
- *Slightly Drowsy:* The driver scans his environment, and his movements look normal (i.e., not slow). He is squinting, but this may be due to the sun. He also seems to be having a conversation with someone, which indicates alertness. He has several head/face touching behaviors, but they occur quickly (if they were slower, it may indicate moderate drowsiness).
- *Moderately Drowsy:* The Driver has several slow eye closures, coupled with periods of a fixed gaze. He appears to have dark circles under his eyes.
- *Very Drowsy:* The driver has sagging facial tone and dark circles under his eyes. He touches/rubs his head at the beginning of the clip, which is a relative indicator of drowsiness. He also opens his window and is drinking coffee, which may be an indication that he is trying to wake himself up. He has several slow eye closures. Slower eye closures or nodding off would have qualified this as an extremely drowsy rating.
- *Extremely Drowsy:* The driver looks somewhat alert in the beginning of the clip, but he appears to be struggling to stay awake after a few seconds. He has slow eye closures (almost looks as though he nods off several times), and appears to have lost some neck control (his head hangs forward). He rubs his face/head, which is a relative indicator of drowsiness. Towards the end of the clip, he does not scan his environment often, and when he does, the movements are large and exaggerated.

The Drowsiness Indicator Checklist

A Drowsiness Behavior and Mannerism Checklist (figure 1) was developed to assist raters in performing ORD ratings. This checklist is to be used as a guide and as a reminder of the various drowsiness indicators that may be observed. This checklist should not be used to directly derive a score, but rather to ensure that all appropriate information is assimilated and referenced in the rating process. Severity and frequency of the different mannerisms need to be considered, some

mannerisms should be weighed more heavily, and all ratings should consider individual differences between drivers. A blank checklist should be used for each new event that gets rated.

ORD Behavior & Mannerism Checklist

Eyes/Eyebrows:	None	Minor	Moderate	Extreme
Rubbing/ Scratching	☐	☐	☐	☐
Blank/Fixed Stare	☐	☐	☐	☐
Squinting	☐	☐	☐	☐
Excessive/Hard Blinking	☐	☐	☐	☐
Slow Closure	☐	☐	☐	☐
Unfocused rolling	☐	☐	☐	☐
Glassy/Glazed	☐	☐	☐	☐
Raise/Open Wide	☐	☐	☐	☐
Lower/Scowl	☐	☐	☐	☐

Body:				
Slumping/Slouching/ Leaning	☐	☐	☐	☐
Sighing	☐	☐	☐	☐
Stretching	☐	☐	☐	☐
Body Rolling/Slack Muscle Tone	☐	☐	☐	☐
Body Position Change (restlessness)	☐	☐	☐	☐

Mouth:	None	Minor	Moderate	Extreme
Yawning	☐	☐	☐	☐
Biting/Licking Lips	☐	☐	☐	☐
Tongue Motion	☐	☐	☐	☐

Face:				
Rubbing/Holding	☐	☐	☐	☐
Facial Contortions	☐	☐	☐	☐
Slack Muscle Tone	☐	☐	☐	☐

Neck/Head:				
Hair: Scratching/ Straightening	☐	☐	☐	☐
Rubbing/Holding	☐	☐	☐	☐
Head Leaning (back or side, unsupported	☐	☐	☐	☐
Head Position Change	☐	☐	☐	☐
Head Nodding/ Drooping	☐	☐	☐	☐

Other Notes:

Figure 1: The ORD Behavior & Mannerism Checklist is a guide to assist the researcher or data reductionist in assigning reliable, reproducible ORD ratings. A blank checklist should be used for each new rating.

Instructions for Determining and Recording ORD Ratings

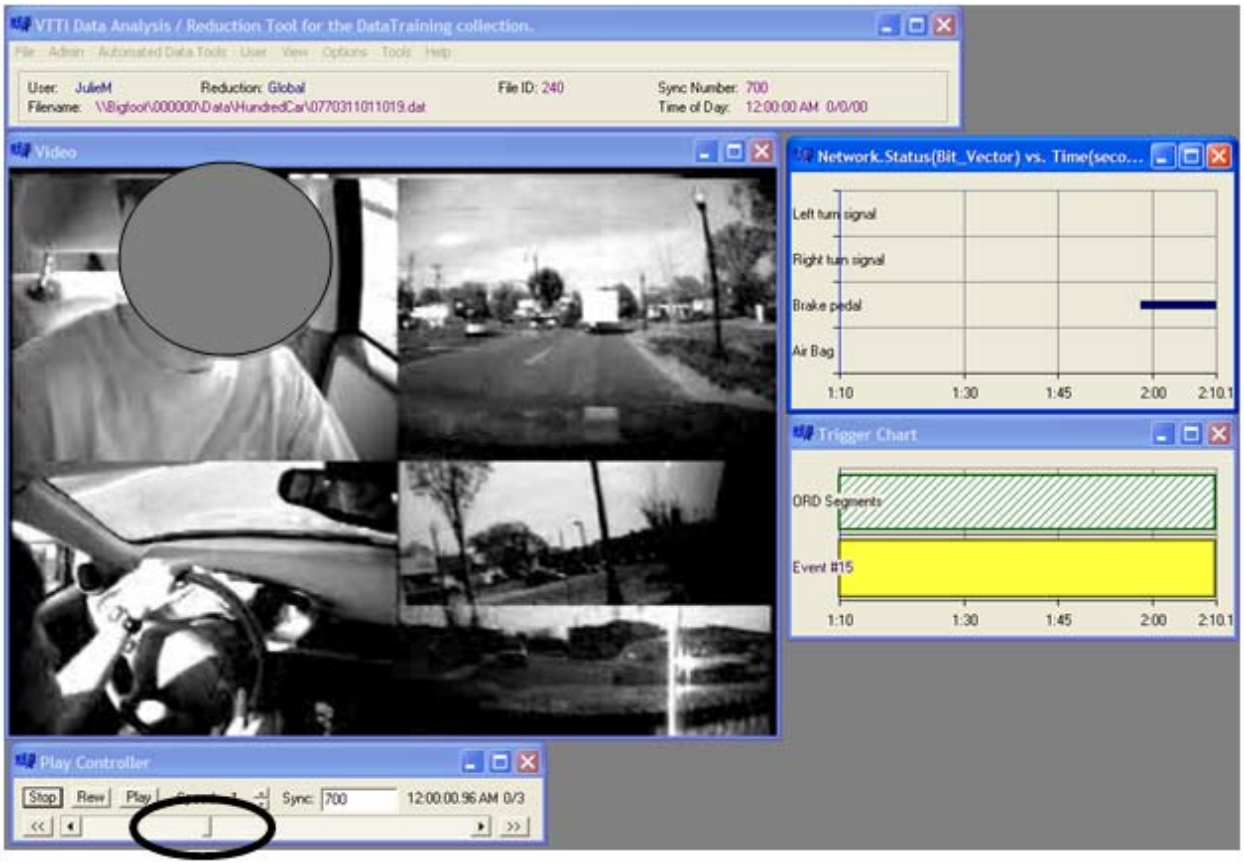
1) Open DART & log into the Data Training collection, Global reduction

2) Open Query Tool & load the ORD Test Events query. Run the query.

3) Refer to the Event list provided to you during training. You are required to complete the events in the order listed for statistical purposes. Each rater is given a different order. Locate the first event on your list by referring to the Event_ID in Query Tool. Double click on that event to load it into DART.

4) Open the following Views:
    a. Video & Play Controller

b. Triggers
c. Network Status



5) Open the event rows of the Trigger chart by right clicking in the chart and selecting "Validation". Once this is done, you will see the yellow event similar to what is shown above.

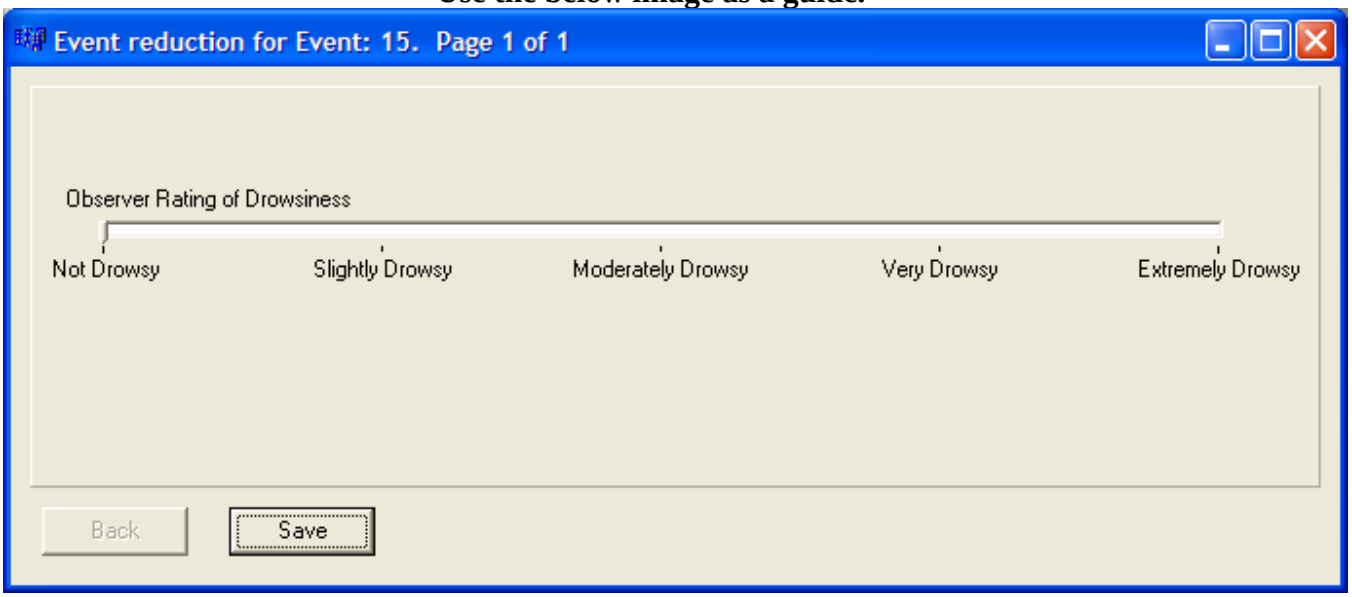
6) Locate the event that you need to complete (the Event_ID is provided on the Y-axis of the Trigger chart), and place the cursor at the start of that event.

7) On the scroll bar in the play controller, click to the left of the cursor (circled above) 6 times. This will take you back from the event exactly 60 seconds. You will analyze 60 seconds of video before the event starts.

8) Right click on the event in the Trigger Chart, and select the "ORD Training" question reduction. This will open the slider tool for you to make your rating.

9) Resize the slider tool that opens up. The slider tool should be 7 inches wide. Use a ruler, or use the image below as a guide. (Note, if you complete all events in one session, you will only need to resize this window once.)

Your ORD rating window should be exactly 7 inches wide on your monitor.
Use the below image as a guide.



10) Watch the video. You may watch the video as many times as you feel you need. Also, check off behaviors and mannerisms on your ORD Behavior and Mannerism Checklist.

11) To make a rating, consider the video you've watched and the notes you've made on the Checklist, and refer to the scale definitions located on page 2 of this manual.

12) Make your rating by sliding the marker (located at "Not Drowsy" to start) across the rating scale and dropping it at the desired point. You can adjust the position of the marker as many times as needed.

13) When satisfied with your rating, click the "Save" button.

14) Return to Query Tool and locate the next event on your list. Repeat steps 6-13 until all events have been completed.

APPENDIX B: EXPERT RATING SCORES AND GOLD STANDARD SCORES

Video Segment	Expert 1	Expert 2	Expert 3	Average (Gold Standard)	Max	Min	Range
1	100	100	97.9	99.3	100	97.9	2.1
2	46.6	53.1	51.3	50.3	53.1	46.6	6.5
3	95.1	81	89.4	88.5	95.1	81	14.1
4	98.5	98.7	82.1	93.1	98.7	82.1	16.6
5	14	1.8	15.4	10.4	15.4	1.8	13.6
6	38.3	24.9	15.8	26.3	38.3	15.8	22.5
7	18	29.5	43.1	30.2	43.1	18	25.1
8	67.3	71.3	47.8	62.1	71.3	47.8	23.5
9	76.2	50.9	63.6	63.6	76.2	50.9	25.3
10	42.6	53.7	43.6	46.6	53.7	42.6	11.1
11	43.7	38.2	49.2	43.7	49.2	38.2	11
12	53.6	62.5	64.5	60.2	64.5	53.6	10.9
13	80.2	61.6	87.5	76.4	87.5	61.6	25.9
14	71.5	64.8	63.9	66.7	71.5	63.9	7.6
15	84.4	88.1	84.6	85.7	88.1	84.4	3.7
16	76.9	69.2	76	74.0	76.9	69.2	7.7
17	24.4	30.3	13.1	22.6	30.3	13.1	17.2
18	12.9	1	22.4	12.1	22.4	1	21.4
19	16.6	21.7	16.3	18.2	21.7	16.3	5.4
20	7.6	12.4	6.9	9.0	12.4	6.9	5.5
21	6.8	13.6	10	10.1	13.6	6.8	6.8
22	25.9	50	37.5	37.8	50	25.9	24.1
23	71.4	53.6	51	58.7	71.4	51	20.4
24	89.7	62.5	62.4	71.5	89.7	62.4	27.3
25	46.6	53.1	51.3	50.3	53.1	46.6	6.5
26	14	1.8	15.4	10.4	15.4	1.8	13.6
27	25.9	50	37.5	37.8	50	25.9	24.1
					Avg Range = 14.80		

APPENDIX C: PROTOCOL FOR ESTABLISHING ORD RATER PROFICIENCY

Establishing Proficiency

When training new ORD raters, it is important to establish a level of proficiency through post-training testing, and then to maintain that proficiency by re-testing and, if necessary, re-training at regular and frequent intervals (e.g., every 1-2 weeks). The below protocol represents the recommended practices for establishing and maintaining rater proficiency.

1. **Train:** Conduct a formal training session using the ORD Training Protocol. In this step, the material in this protocol (including all video examples) should be reviewed with the ORD trainees, preferably as a group so that all raters hear and see the same materials and to permit questions and discussions. This training session is estimated to be 2 hours long, and may be conducted as one session, or broken into two 1-hour sessions.

   ORD raters should be encouraged to retain and refer back to this protocol (including category descriptions, video examples, etc.) both during the test, and when performing actual ORD ratings.

2. **Test:** Once the training is complete, all trainees should complete the first proficiency test. This test is comprised of 24 unique 1-minute video clips representing the full range of drowsiness levels and a variety of drivers of different facial features, skin color, etc. Three of these video clips should be repeated during the test (for 27 total events on the test) in order to get an intra-rater consistency measure. Videos should be presented either in a random order to each rater, or following some other balanced ordering scheme (e.g., Balanced Latin Square).

   Raters should be provided with enough ORD scales and the ORD Behavior and Mannerism Checklists so that clean entry forms are used for each event. Training should be provided in the use of the equipment used (e.g., computer software, printed checklists, etc.). Practice events may be provided at the reduction manager's discretion, but only if all raters being tested complete the same practice events. No feedback is given on the practice events.

   During the test, calibration events for each driver may or may not be viewed, at the discretion of the reduction manager; however, this decision should be consistent across all raters taking the test. If calibration events are not viewed, it is anticipated that this test will take a trained rater approximately 1.5 hours to complete. Note that the expected variance in ORD scores is slightly higher with the absence of calibration events, and this should be taken into consideration when scoring the test. If calibration events are viewed for each driver, then these calibration events should be pre-selected, and the same calibration events should be viewed by all raters being tested. Calibration events will

significantly increase the time required to take the test, depending on how many events are selected for each driver. At least three events per driver are suggested if calibration events are used. NOTE: calibration events should always be used during actual ORD sessions when new drivers are started. They are optional only during proficiency tests.

3. **Evaluate:** Once all raters have completed the proficiency test, the scores should be converted to numerical equivalents using the 0-100 scale illustrated in the training protocol. Individual ORD scores should then be compared to other raters taking the test and to the average of at least three "gold standard" rater scores. (The "gold standard" raters should be highly experienced and tested ORD raters.) If the range across raters is more than 30 points, or if any individual's rating falls more than 30 points (absolute difference) from the average of the gold standard average, those raters should be re-trained and re-tested (with a new set of events). If accuracy standards are not met after a second proficiency test, the reduction manager may decide to remove that rater from the ORD task and train someone else.

4. **Re-Test:** Re-testing is necessary either when trained raters fail an initial proficiency test or when the ORD task is conducted over a period of time. In the latter case, it is suggested to re-test all raters at least every two weeks to ensure that accuracy is maintained over time. It is also recommended to conduct shortened "refresher" training sessions at regular intervals, or at minimum when/if proficiency re-test results begin to decline. Events present on re-tests should be different from those present in earlier tests, except for a small sample (e.g., 3 events) that should be repeated from test to test in order to monitor intra-rater and test-retest reliability. Accordingly, each re-test should have 27 total events to be rated: 21 new events, plus 3 events repeated from a previous test, plus 3 events from the new group that are repeated during the same test.

REFERENCES

(1) Wierwille, W. W. & Ellsworth, L. A. (1994). Evaluation of driver drowsiness by trained observers. *Accident Analysis and Prevention*, *26*(5), 571-581.

(2) National Sleep Foundation (2008). 2008 Sleep in America Poll. Retrieved February 2, 2009 from http://www.sleepfoundation.org/atf/cf/%7Bf6bf2668-a1b4-4fe8-8d1a-a5d39340d9cb%7D/2008%20POLL%20SOF.PDF

(3) Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., Perez, M. A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z. R., & Jermeland, J. (2006). The 100-car naturalistic driving study; Phase II- Results of the 100-car field experiment. Contract No. DTNH22-00-C-07007 (Task Order No. 06). Blacksburg, VA: Virginia Tech Transportation Institute.

(4) McCartt, A. T., Rohrbaugh, J. W., Hammer, M. C., & Fuller, S. Z. (2000). Factors associated with falling asleep at the wheel among long-distance truck drivers. *Accident Analysis and Prevention*, *32*, 493-504.

(5) Transportation Safety Board (1990). Fatigue, alcohol, other drugs, and medical factors in fatal-to-the-driver heavy truck crashes. Retrieved November 29, 2006 from http://www.ntsb.gov/publictn/1990/SS9001.htm

(6) Hanowski, R. J., Wierwille, W. W., Garness, S. A., & Dingus, T. A. (September, 2000). Impact of local/short haul operations on driver fatigue, final report. Report No. DOT-MC-00-203. Washington, DC: U.S. Department of Transportation, Federal Motor Carrier Safety Administration.

(7) Wiegand, D. M., Hanowski, R. J., Olson, R., & Melvin, W. (2008). *Fatigue analyses from 16 months of naturalistic commercial motor vehicle driving data*. Blacksburg, VA: National Surface Transportation Safety Center for Excellence.

(8) Emery, F. E., & Trist, E. (1960). Socio-technical systems. In C. W. Churchman and M. Verhulst (Eds.) *Management Sciences Models and Techniques*, Vol. 2 (pp. 83-97). London: Pergamon Press.

(9) Hanowski, R. J., Blanco, M., Nakata, A., Hickman, J. S., Schaudt, W. A., Fumero, M. C., Olson, R. L., Jermeland, J., Greening, M., Holbrook, G. T., Knipling, R. R., & Madison, P. (2005). The Drowsy Driver Warning System Field Operational Test: Data collection methods final report. Contract No. DTNH22-00-C-07007. Washington, DC: National Highway Traffic Safety Administration.

(10) Hickman, J. S., Knipling, R. R., Olson, R. L., Fumero, M. C., Blanco, M., & Hanowski, R. J. (2005). Heavy vehicle-light vehicle interaction data collection and countermeasure research project: Phase I – Preliminary analysis of data collected in the Drowsy Driver Warning System Field Operational Test. DTNH22-00-C-07007, Task Order #21. Federal Motor Carrier Safety Administration: Washington, DC.

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

(12) Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design and analysis: An integrated approach. New Jersey: Lawrence Erlbaum Associates, Inc.