

# **Practical methods of ORD measurement and assessor training**

**13. Oct. 2025**

**IWG-DDADWS #9 Workshop**

**JAMA Vehicle Safety Committee  
Technical Working Group of Driver monitoring**

# Objectives of the presentation

## Objectives:

- To introduce some practical examples of the ORD, which evaluates "how drowsy" a driver feels during driving based on facial expressions, and to present the applicability of ORD as a drowsiness measurement method for DDAW regulations.

## Approach:

- To ensure that drowsiness measurement is accurate and consistent regardless of when, who, or how many times they are performed, it should show intersubjectivity and reliability.
  1. ORD utilizes humans' natural ability to recognize facial expressions.
  2. Although ORD is a subjective method performed by third parties, measurement by multiple assessors can enhance "intersubjectivity".
  3. Present some assessment procedures and practical examples of training protocols to ensure the "reliability" (consistency and stability) of ORD assessment values.

# Today's Presentation agenda

1. Relationship between drowsiness and facial expressions
2. Practical examples of ORD (Observer Rating of Drowsiness)
  - Intersubjectivity
  - Kitajima's method as the most popular method in Japan.
  - Practical examples of a training to reduce variability between experts and beginners
3. Summary and next step

# Relationship between drowsiness and facial expressions

# Drowsiness shows on the face

上眼瞼筋によって目が大きく見開かれると、表情が強烈になる。しかし、リラックスした表情にも上眼瞼筋の伸び縮みが影響している。



## A. びっくり目

上下のまぶたを思いきり開いている。上眼瞼筋が最大限収縮した状態で、黒目の上にはたいてい少し白目が見える。まぶたの描くアーチはより大きくなり、じっと見つめる感じになる。



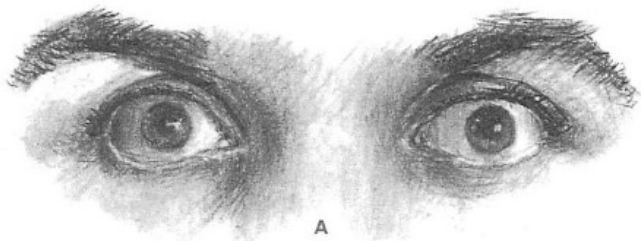
## B. 警戒した目つき

自然な状態の目(C)に比べて、やや大きく見開いており、黒目が少し上まぶたで覆われている(上図)。

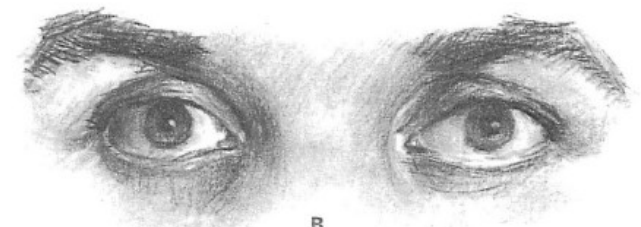


## C. 自然な状態の目

黒目の上まぶたがかかっているが、うんと上の方なので半分以上は見えている。



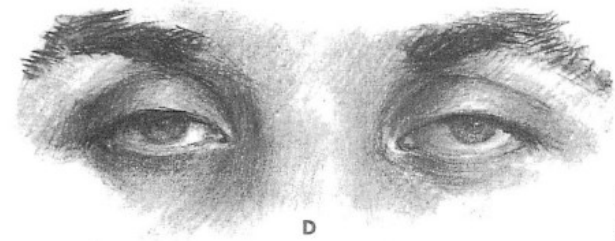
A



B



C

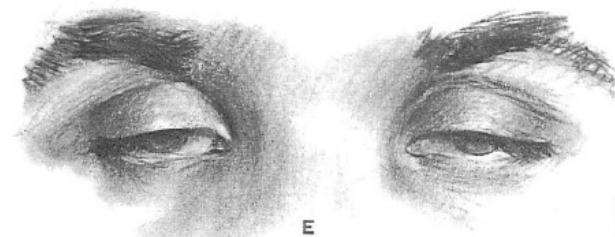


D



## D. 眠そう目

黒目がちょうど半分隠れている。上眼瞼筋が緩んでまぶたが下がり、瞳孔も部分的に覆われている。このような目は一時的なもので、意識して長時間保つことはできない。

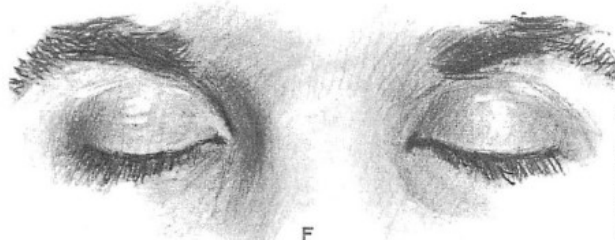


E



## E. 半分意識のない目

目覚めればかりか、眠りに落ちる直前の瞬間的な目。瞳孔の大部分が覆われ、ほとんど見えていない状態。黒目の3分の2が隠れている。



F

## F. 閉じた目

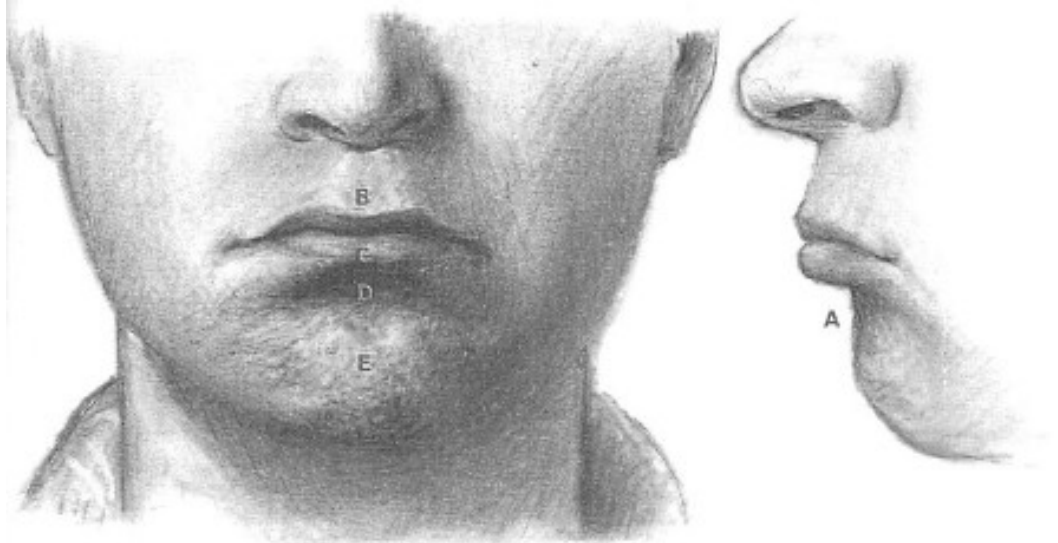
上眼瞼筋が完全に伸びて、目が閉じられた状態。下まぶたと上まぶたがくっつき、上まつ毛と下まつ毛が重なって強調されている。顔のデッサンではまつ毛を描かないことが多いが、この場合、まつ毛を描くことが必要である。

Source: Gary Faigin, The Artist's Complete Guide to Facial Expression(1990)

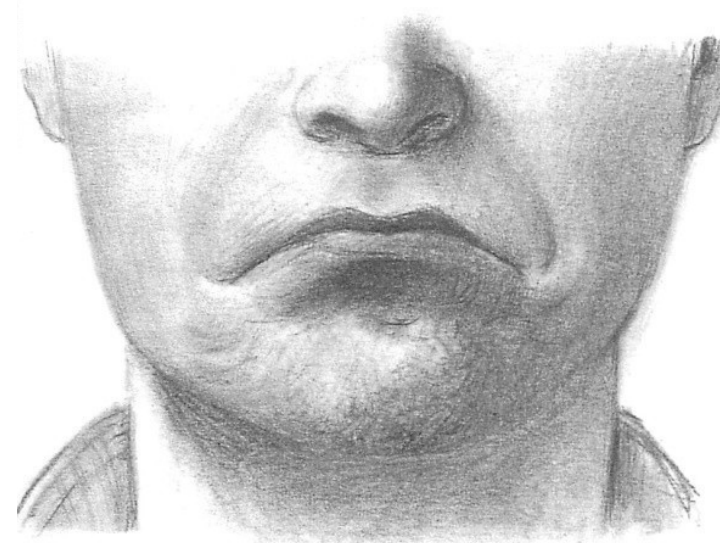
**Drowsiness is most evident in the eyes.  
It is important to continuously observe the area around the eyes.**

# Drowsiness shows on the face

すねた表情を作る筋肉



オトガイ筋の基本動作はあごの皮膚を歯の根元の方に引っ張ることだが、この時、あごの皮膚は平らになる。唇は持ち上げて突き出る（A）。持ち上げる力がもっとも強いのは中央部だが、唇は全体的に薄くなる（B）。唇の合わせ目はまっすぐになり、下唇の縁はくっきり明るくなる（C）。中央部が持ち上がるため、口角は引き下げられる。唇の下には濃い影がで、くぼみはリラックスした口よりも際立つ（D）。隆起した唇のような形の「島」があごで、あばせができる（E）。



顔をすぼめる

三角筋とオトガイ筋は連動して収縮し、顔をすぼめる。ちょうど肩をすくめるような感じだ。その動きがかすかならずねた表情になる。三角筋の動きによる下がった口角と、釣り針形の深いしわ、オトガイ筋の動きによる平らな唇と、ひだのできたあごが特徴だ。下唇が上唇より多くの部分を占め、突き出して、大きく見える。

Source: Gary Faigin, The Artist's Complete Guide to Facial Expression(1990)

Not only the area around the eyes but also the entire facial expression, including the mouth, should be observed.

# Relationship between ORD level and facial features

ORD level	Drowsiness description	Typical observed behavior (from Uchiyama's paper)
D1	Not drowsy	Fast and frequent eye movements, Frequent body movements, Eye blinks with fast eyelid movements, Regular time intervals between eye blinks, Active body movements
D2	Slightly drowsy	Slow saccadic eye movements, Lips open, Infrequent eye movements, Drooping eyelids
D3	Moderately drowsy	Frequent slow eye blinks, Mouth movement, Sitting position change, Touching face, Frequent eye blinks, Less than half obscured pupils, Yawning, Tired-complexion
D4	Very drowsy	Voluntary eye blinks, Frequent yawning and deep breathing, Unnecessary body movements such as shaking head and up and down, movement of the shoulders, etc. Slow blinking and SEM, Staring blankly at a single point, Inability to focus, More than half obscured pupils
D5	Extremely drowsy	Closed eyes, Forward tilted head, Backward tilted head, Sagging cheeks
S	Sleeping	D5 continues and does not awaken. The state is judged as "likely to be asleep".

Source: Uchiyama et al., Convergent Validity of Video-based Observer Rating of Drowsiness (ORD), against Subjective, Behavioral, and Physiological Measures(2022)

## Practical Examples of ORD

**-Methods and training to enhance the consistency and stability of assessment values-**

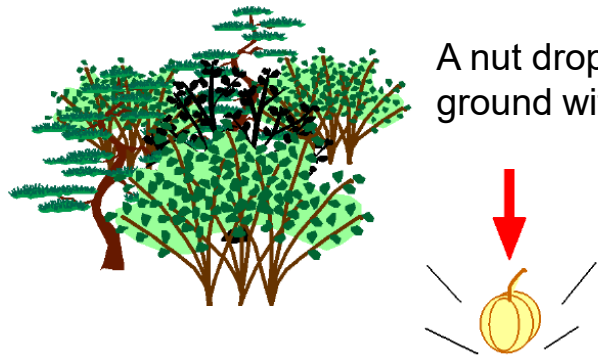
\*Explanation of the Kitajima's method, which is commonly used in Japan

# Importance of Intersubjectivity

- Intersubjectivity: A state in which multiple people can share and understand subjective experiences.
- High intersubjectivity leads to higher reproducibility and reliability, approaching objectivity.
- However, it is known that objectivity is not complete and depends on the context, culture, and observers.

Example)

In the forest..



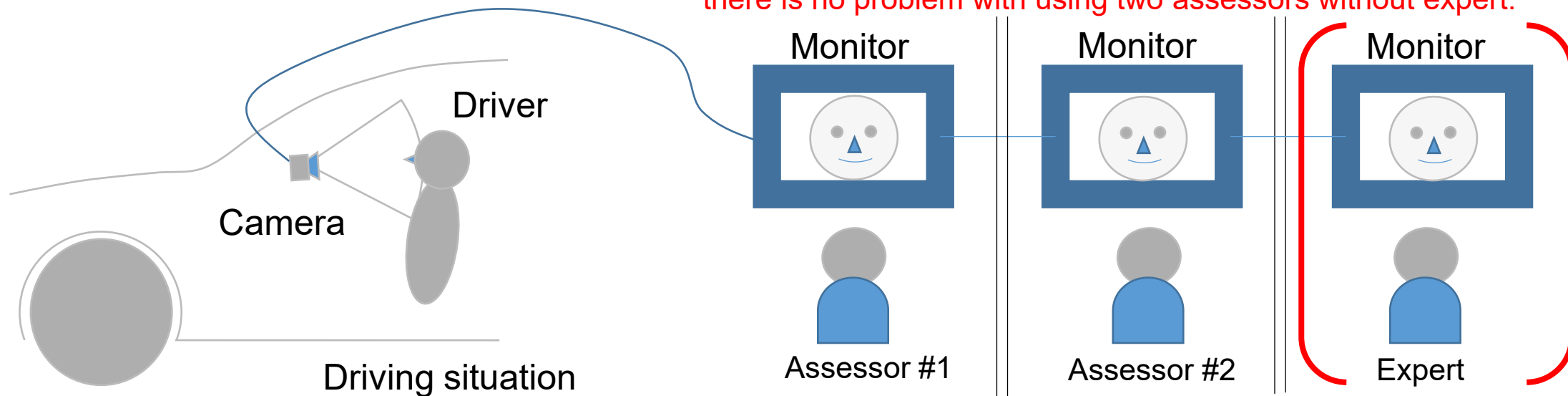
People reported the same content in verbal.

Multiple people could share subjective experience.

->Corresponding Intersubjectivity

**To enhance the reliability of ORD, it is important to increase "intersubjectivity."**

# Example of Kitajima's Method



- ✓ **Two assessors independently observe** and evaluate the facial expressions of drivers in video.
- ✓ **Concordance rate between the two assessors and the expert** is calculated, and agreement above the standard is confirmed\*. (\*Concordance rate  $\geq 0.7$  is OK.)
- ✓ The average of the two scores is used as the representative drowsiness value.

## Concordance rate

$$\frac{\sum_{i=1}^n \left[ 1 - \frac{Estim. 1_i - Estim. 2_i}{Dmax_i} \right]}{n}$$

*Estim.1*: evaluated value of assessor1  
*Estim.2*: evaluated value of assessor2  
*Dmax*: Max gap of evaluated value  
*n*: num of data

# Example of Kitajima's Method

## Facial expression

**Very high alert**

**Moderate**

**Very low alert**

**Drowsy**

**1**

**2**

**3**

**4**

**5**

**S**

**Time**

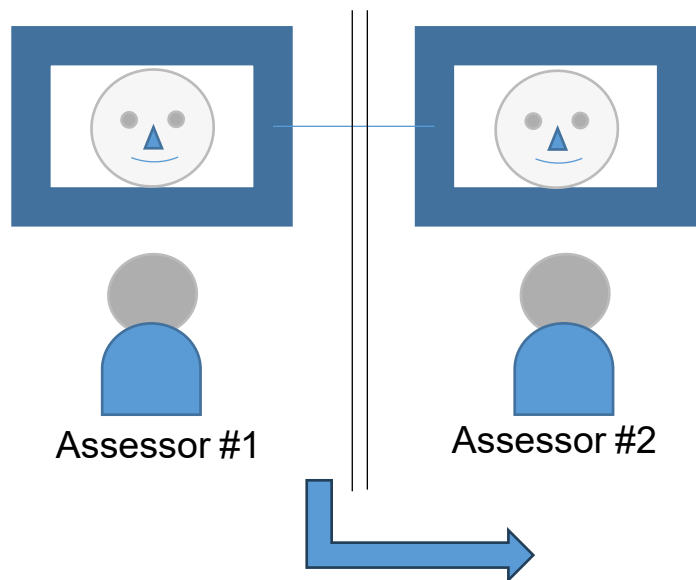
Source: Gary Faigin, The Artist's Complete Guide to Facial Expression(1990)

- ✓ Assessment is conducted every 5 seconds.
- ✓ Assessment is conducted using a 5-point(+1) equal interval scale.
- ✓ The entire facial expression is observed, focusing on the eyes and mouth.
- ✓ Drowsiness is judged based on changes in facial expressions over time.

Drowsiness level (incl. sleeping)	
D1	Not drowsy
D2	Slightly drowsy
D3	Moderately drowsy
D4	Very drowsy
D5	Extremely drowsy
S	Sleeping

# Example of Kitajima's Method

To ensure the reliability of ORD, suppressing variability between assessors if there is a scoring bias.



When either Assessor #1 or Assessor #2 has a bias in their scores, add a **weighting factor until the concordance rate exceeds 0.8**, then average the two assessors.

## Weighted Cohen's Kappa

$$p_o = \sum_{i,j} w_{ij} O_{ij} \quad \text{: Observed agreement}$$

*O<sub>ij</sub>*: Observed proportion in the contingency table

$$p_e = \sum_{i,j} w_{ij} E_{ij} \quad \text{: Expected agreement by chance}$$

*E<sub>ij</sub>*: Expected proportion

$$w_{ij} = \left( \frac{i - j}{k - 1} \right)^2 \quad \text{: Quadratic weights}$$

$$\kappa_w = \frac{p_o - p_e}{1 - p_e} \quad \text{: Weighted Kappa}$$

Reference:

[Cohen's kappa - Wikipedia](#)

**Variability between assessors can be suppressed by using concordance rate and “weighted kappa”.**

## Summary of ORD Reliability (=Intersubjectivity)

- ORD utilizes humans' advanced facial expression recognition abilities.
- Although ORD is a subjective assessment method by third parties, the Kitajima's method ensures intersubjectivity by keeping the low variability between independent assessors within a certain criterion (consistency and stability can be quantitatively demonstrated).
- **Training of assessors** is important to the improve concordance rates.

The next section introduces examples of assessor training.

# Practical Examples of ORD Training

# Practical example of training in one Japanese OEM



1.-  
1

## Step1: Establishing a 5-point scale for drowsiness of observed subject

Before the assessment, play various scenes from the training videos and observe the facial expressions at each moment to determine which expressions correspond to which numerical ratings.

1.- 2.-  
2 2

## Step2: Play the video and conduct the assessment at 5-second intervals.

Based on the results from step1, try to begin the evaluation.

If you find the difficulty, it indicates that the “5-level drowsiness criteria” have not been sufficiently established, so return to step1.

1.- 2.-  
3 3

## Step3: Review and feedback of the assessment results

Compare the assessment results with the expert’s one and score them (considering concordance rate and kappa coefficient).

If there is a significant discrepancy between the trainee’s and the expert’s, provide feedback on the specific segments and the presumed causes.

**It is important to establish a 5-point scale of facial expression assessment**

# Real-time training & Confirmation test



## Step1: Establishing criteria for the 5-point (plus $\alpha$ ) facial expression assessment

Before the assessment, play various scenes from the training videos and observe the facial expressions at each moment to determine which expressions correspond to which numerical ratings.

## Step2: Play the video and conduct the assessment at 5-second intervals.

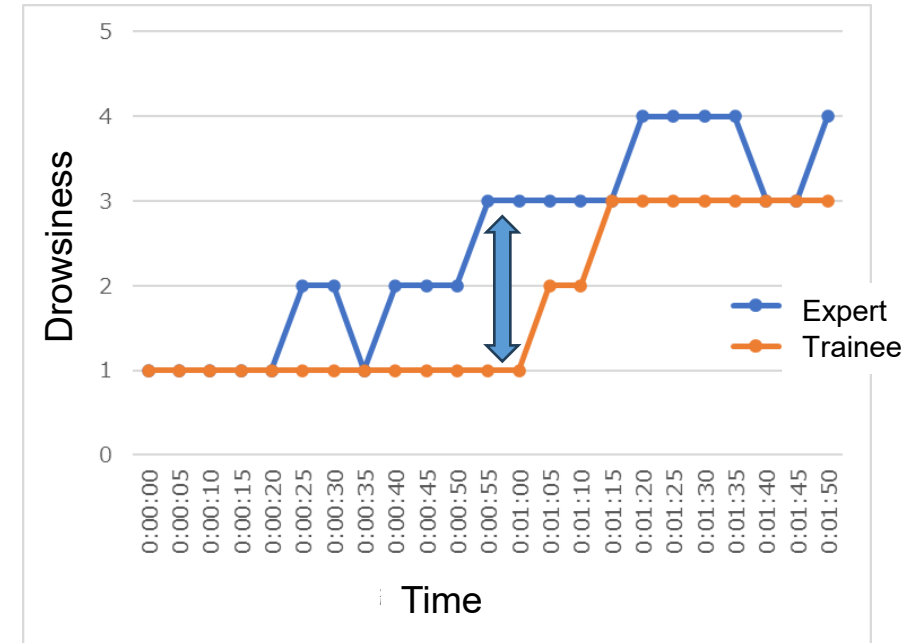
Based on the results from step1, begin the evaluation.

If you find the difficulty, it indicates that the “5-level drowsiness criteria” have not been sufficiently established, so return to step1.

## Step3: Review and feedback of the assessment results

Compare the assessment results with the expert’s one and score them (considering concordance rate and kappa coefficient).

If there is a significant discrepancy between the trainee’s and the expert’s, provide feedback on the specific segments and the presumed causes.



- Observe the driver's facial expressions during driving in real time and assess the drowsiness level every 30 seconds (every 5 seconds during the training) using a 5-point scale.
- Conduct the training in pairs of an expert and a trainee, **calculate the concordance rate, and continue the training until the rate meets the threshold of 0.8.**
- The trainee is given feedback on concordance rate between the expert's results and the trainee's results.
- The expert will provide comments on scenes where there is a large difference between the expert's and the trainee's.

**It is important to provide feedback on this gap with expert assessment**

# Example of Training Protocol at VTTI

A training protocol for ORD using naturalistic driving examples at VTTI

Establish a level of proficiency through post-training testing.

## 1) Train(2h) only

only lectures on 5 levels of drowsiness

➔ Average deviation from expert: **24pt/100**

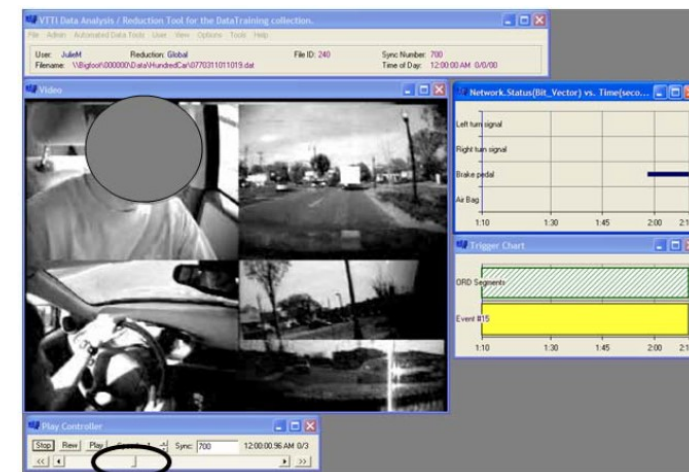
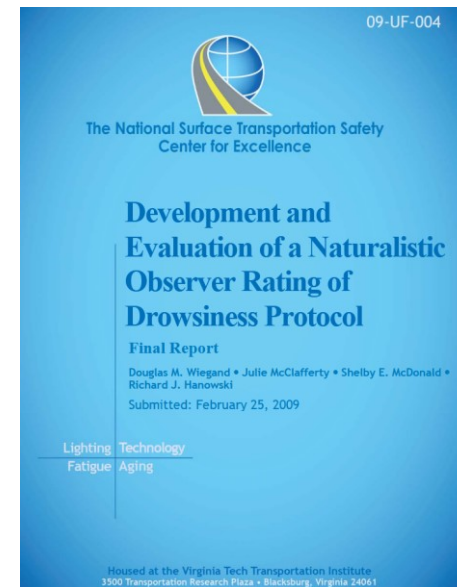
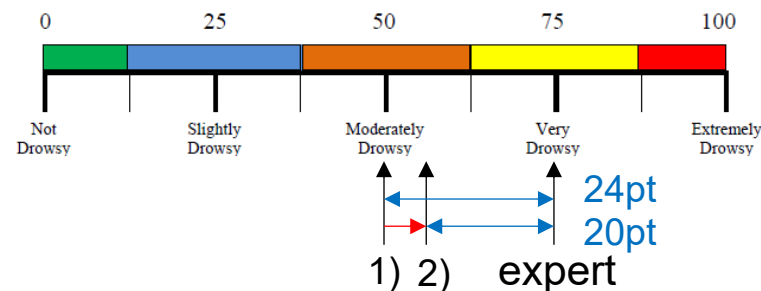
## 2) Train(2h) + Test(1.5h) + Evaluate(1h)

1) + tested and assessed according to a developed training protocol

➔ Average deviation from expert: **20pt/100**

➔ Not only training but combination to test and evaluate session can **gain 4pt/100.**

The report states that repeated training using **videos previously rated by ORD experts** can improve agreement rates with those experts. ➔ **Self-training** is available as well.



Source: [ORD Final Report 022509](#)

**There are also methods for training using videos that have been rated by ORD experts.**

# Assessment and Training Protocols (Comparison Table)

		Kitajima's method	JP's OEM method	VTTI's method
Assessment	Scale	5 points	Same as Kitajima's	0-100 points (of 5 levels)
	Assessment Interval (DDAW: ~5 min)	5 sec	5-30 sec	60 sec
	Num of assessor (DDAW: 3 assessors)	2 (independent) +1 (expert)	2 (independent) +1 (expert)	3 (independent)
	Record keeping	Every 5sec on the sheet	Every 5sec on the sheet (avg. every 5-30 sec)	Every 60sec on the ORD checklist
Training	Training time	(Not disclosed)	~4 days (8h/day)	Min. 2h + 2.5h additional
	Feedback after training (DDAW: from expert)	Compared with expert results	Compared with expert (trained by Kitajima) results	Compared with average of 3 experts
	Completion criteria (DDAW: concordance rate ≥0.7)	Concordance rate ≥0.7	Concordance rate ≥0.8	Point difference with experts <±30

≡ concordance rate>0.7

**Although there are some differences among those methods, the concepts are the same.**

## Summary of ORD training

- Training is conducted by pairing an expert and a trainee, comparing their assessment values, and repeating feedback on the differences.
- Finally, trainees are recognized as assessors when their concordance rate with experts meets a certain criteria (e.g.,  $>0.7$ ).
- Various training methods exist (using videos rated by experts, real-time videos), but the outcomes are considered consistent across methods.

## Summary and Next step

# Summary and Next Step

## <Summary>

To ensure that assessments are accurate and consistent regardless of how many times they are performed, it is necessary to ensure (1) intersubjectivity and (2) reliability of assessment values.

(1) Although ORD is a subjective assessment method by third parties, it possesses "[intersubjectivity](#)".

(2)-1: Through proper training and assessment methods, ORD can ensure that "[expert assessment values & variability between assessors](#)" are within a certain range (reliability can be quantitatively demonstrated).

(2)-2: The "[training protocol](#)" and the "[expert rating criteria](#)" for calculating the concordance rate are approximately standardized, so even beginners (anyone) can make reliable assessment through the proper training.

## <Next Step>

When trying to put ORD into practice in various places, it is necessary to "[developing experts](#)" of assessment.

Feedback from experts helps people master the skills faster and improves concordance rates.

Thank you for listening.

